

版权注意事项：

- 1、书籍版权归作者和出版社所有
- 2、本PDF仅限用于个人获取知识，进行私底下的知识交流
- 3、PDF获得者不得在互联网上以任何目的进行传播
- 4、如觉得书籍内容很赞，请购买正版实体书，支持作者
- 5、请于下载PDF后24小时内删除本PDF。

“

统计学学习地图与计算机应用

”

数据挖掘，大数据，网络存储，数据库管理，人工智能，机器学习，模式识别，数据可视化等技术的基础学科。



大话

统计学

陈文贤 陈静枝 著

清华大学出版社



陈文贤

美国加州大学伯克利分校工业工程博士，曾任台大信息管理系教授兼系主任、国雪城Syracuse大学客座教授、澳大利亚悉尼科技大学UTS客座教授，现任德明财经科技大学信息管理系特聘教授。

陈静枝

美国德州大学达拉斯分校管理科学博士，曾任台大信息管理系教授兼系主任、美国贝尔北方实验室研究顾问、美国国家手工具公司计算机顾问系统分析师，现任台大信息管理系教授。





45000	30000
56000	22456
	103456
	56000
56000	
45000	

大话 统计学



陈文贤 陈静枝 著

清华大学出版社
北京

内 容 简 介

“统计学”是兼具“数学计算”与“图形显示”的课程,所有的统计软件(如 SPSS),并非计算机辅助教学(CAI),因为它们并非“教你学会统计”,而是应该在“学会了统计”以后,再来用它。本书就是这样一本让你从零开始接触统计学,并将其真正应用到工作中的一本书,稳步跟进大数据时代。

本书前后连贯,各章之间也是先后呼应。例如:从概率到抽样,从描述到推断,从检验到因果;每章也是连贯的,开头有引言、观念图,结尾有流程图、思维导图;书中有许多阶层图、分类图、关联图、步骤图、流程图,以及因果表、比较表、决策法则表等。

本书专门的配套软件(中文统计)是在 Excel (2003~2016 版本适用)环境下,安装一个“加载项”,输入统计资料,就可以得到统计结果。“中文统计”可以公开下载,仅提供给合法取得本书之读者使用。

本书适合所有想掌握统计学的读者,也可以作为高校教材,由于内容比较多,教师可自行选择教学内容。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

大话统计学/ 陈文贤, 陈静枝著. — 北京: 清华大学出版社, 2016
ISBN 978-7-302-45018-4

I. ①大… II. ①陈… ②陈… III. ①统计学—基本知识 IV. ①C8

中国版本图书馆 CIP 数据核字(2016)第 218501 号

责任编辑: 栾大成

装帧设计: 杨玉芳

责任校对: 徐俊伟

责任印制: 何 芊

出版发行: 清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址: 北京清华大学学研大厦 A 座 邮 编: 100084

社 总 机: 010-62770175 邮 购: 010-62786544

投稿与读者服务: 010-62776969, c-service@tup.tsinghua.edu.cn

质量反馈: 010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者: 清华大学印刷厂

经 销: 全国新华书店

开 本: 188mm×260mm

印 张: 26.5

字 数: 535 千字

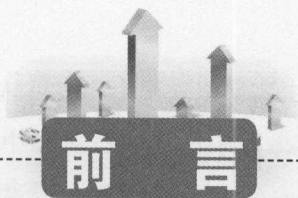
版 次: 2016 年 12 月第 1 版

印 次: 2016 年 12 月第 1 次印刷

印 数: 1~4000

定 价: 59.00 元

产品编号: 070745-01



百闻不如一见，兵难隄度，臣愿驰至金城，图上方略。

——《汉书·赵充国传》

因为计算机的普及，管理学院和工学院有些课程，在教材编写与教学方法方面，有结构性的改变。尤其是有关数学计算或图形显示的课程。这些课程应该附有应用软件的使用，配合教学。

“统计学”就是有“数学计算”与“图形显示”的课程。本书作者 1990 年在中国台湾出版的《统计学》一书，就附有“中文统计”软件，以前是在 DOS 操作系统下操作，现在则是在 Excel 上安装“加载项”操作。“中文统计”软件仅被提供给教师和购买本书的读者使用。

所有的统计软件（包括 SPSS），并非计算机辅助教学（CAI），因为它们并非“教你学会统计”。而是应该在“学会了统计”以后，再来用它。有了统计软件以后，大多数的统计问题，只要会选择统计功能，输入数据，就可以得到答案。但是使用统计软件要注意下列应用：

（1）能够判断统计数据，应该（适合）用哪一种统计方法（模型）。

（2）了解并解释统计方法计算结果（答案）的意义。

学习要有地图，学习地图或者知识地图，地图告诉你：你在这里（here）、为什么在这里（why）、要往哪里去（where）、如何去（how）、会得到什么（what）。其中 why 是假定条件（为何适用这个方法 how），而 where 是目的，what 是答案。

1980 年，Wonnacott 统计学第一章的开头引用了一句话：“He uses statistics as a drunken man uses lampposts – for support rather than for illumination.”

“人们利用统计，就好像醉汉利用路灯，是为了支撑，而不是照明。”

还有一个醉汉与路灯的故事：一个醉汉在夜晚的路灯下找东西，有个路人问他在找什么，醉汉说：“钱包。”路人就帮他找，两个人找了很久，但就是找不到。路人问：“你确

定是掉在‘这里’吗?”醉汉说:“我不知道掉在‘哪里’。”路人问:“为什么要在‘这里’找?”醉汉说:“因为‘这里’有路灯比较亮。”

利用统计学时,要注意假定条件是否符合,不要削足适履,不要因为“这个”方法,比较熟悉、比较容易用,就要用它来找答案,结果找到的答案根本不对。

本书取名《大话统计学》是清华大学出版社责任编辑栾大成先生的建议,“大话统计学”的意思不是说大话的统计学,而是让读者可以“大声说话”的统计学。因为本书的图画很多,所以本书也是“大画统计学”。

本书特色:

- (1) 本书前后连贯,有前言、总论、结语。结语有:统计问题分类、统计概念复习。
 - (2) 各章之间也是先后呼应。例如:从概率到抽样,从描述到推断,从检验到因果。
 - (3) 每章也是连贯的:开头有引言、观念图,结尾有流程图、思维导图。
 - (4) 书中有许多阶层图、分类图、关联图、步骤图、流程图,以及因果表、比较表、决策法则表等,所以本书希望为统计学的学习地图。
 - (5) “中文统计”软件是在 Excel (2003 ~ 2016 版本适用) 环境下,安装一个“加载项”,输入统计资料、就可以得到统计结果。“中文统计”可以公开下载,仅提供给合法取得本书的读者使用。
 - (6) 中文统计的功能列表,配合本书章节设计。输入原始数据(观测值),可做描述或推断统计的计算。如果没有原始数据,只有样本容量、样本平均数、方差、比例等数据,那么使用“快速估计”或“快速检验”,可以得到推断结果。
 - (7) 为了节省篇幅,每章的部分例题、习题,放在互联网中,以便读者下载。
 - (8) 每章的部分案例讨论也放在互联网中,可提供学生做学期报告。这些案例的数据也可以下载,以方便学生使用。
 - (9) 补充教材如:分组资料描述统计、自力(Bootstrap)估计法、非参数统计补充、多因素方差分析、多元回归等,也放在互联网中。
 - (10) 因为配合计算机程序,包括 Excel 函数和命令,所以,所有的统计公式和计算步骤,都被很清楚地一一列举出来。由于强调应用导向,所以多数公式没有证明。
- 有了地图和交通工具(计算机软件),就可以快速地到达目的地。但是,如果一路直达目的地,就会错过沿途美妙的风光。所以,初学者还是要先走过一趟(自己计算了解过程),再利用交通工具(计算机),检查结果的正确性。

本书内容很多,所以教师可自行选择教学内容。感谢协助本书编辑及计算机软件的中国台湾大学资管所研究生。

感谢协助编撰本书及计算机软件的台大资管所研究生，使本书能够顺利出版。

由于作者的水平有限，本书中难免有许多错误和疏漏之处，恳请各位学者专家和读者，提出批评和建议，以便进一步修订和改进。

陈文贤 陈静枝 谨识

2016 年 9 月于台北

资源下载

1. 微云下载

- 例题
- Excel 文档
- 习题
- 补充教材
- “中文统计” 安装程序（随时更新）

<http://share.weiyun.com/6989e41582e4c3018d11972acb5b7392>



2. 百度云下载教学资源（要输入百度账号登录）

- PPT
- 习题解答

<http://pan.baidu.com/disk/home#list/path=%2F>



“中文统计” 安装说明

(1) 请到清华大学出版社 www.tup.com.cn 或 cnstat.weebly.com 下载“中文统计.exe”。

(2) 先执行“中文统计”安装程序。

然后，单击“Next >” → “Next >” → “Install” → “Finish”，完成“中文统计”安装，如图 A-1 和图 A-2 所示。

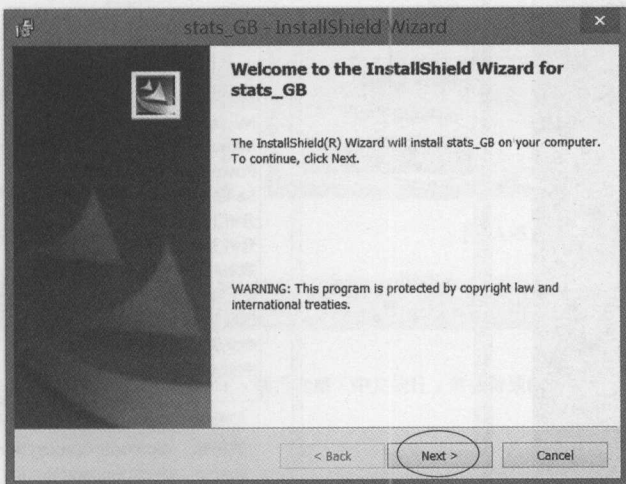


图 A-1 安装操作 1

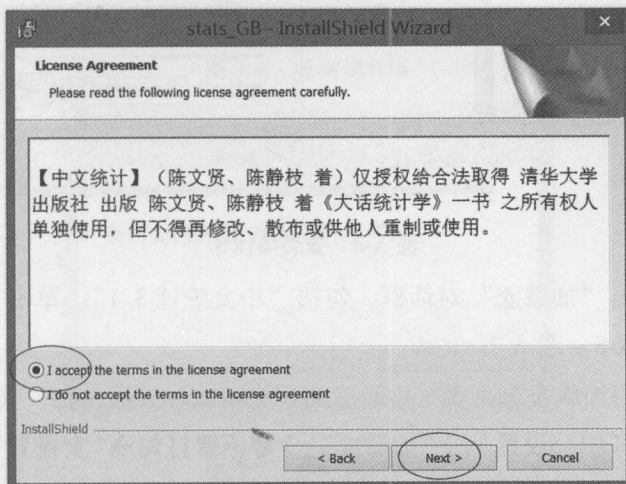


图 A-2 安装操作 2

(3) 请打开 Microsoft Excel (2013, 2016), Excel 2010 以下版本也可适用, 界面不同, 操作如图 A-3 所示。

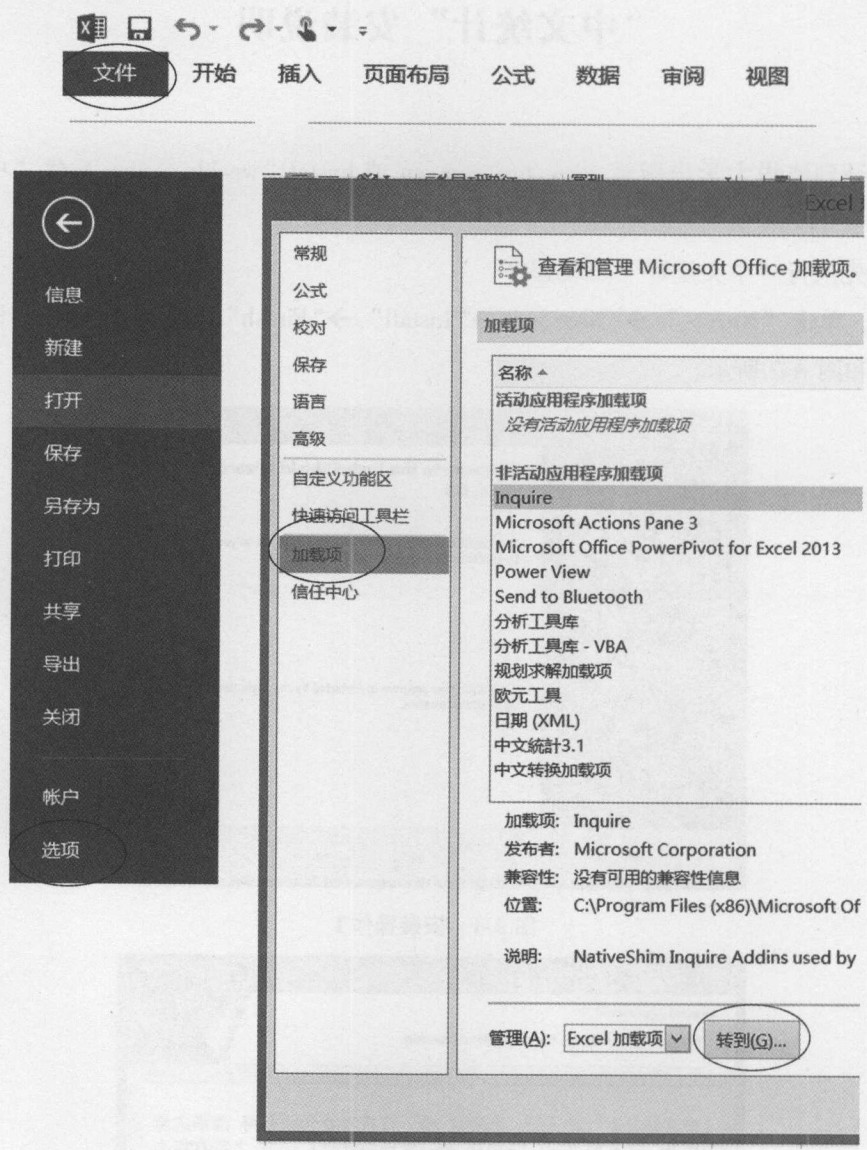


图 A-3 安装操作 3

(4) 然后, 弹出“加载宏”对话框, 勾选“中文统计 3.1”, 单击“确定”, 如图 A-4 所示, 然后出现图 A-5 ~ 图 A-7。

(5) 如果“可用加载宏”列表框中没有出现中文统计, 那么在“文件”→“选项”→“信任中心”→“信任中心设置”→“加载项”→都不要打勾→“宏设置”→单击“启用所有宏”。关闭 Excel, 再重新打开 Excel。

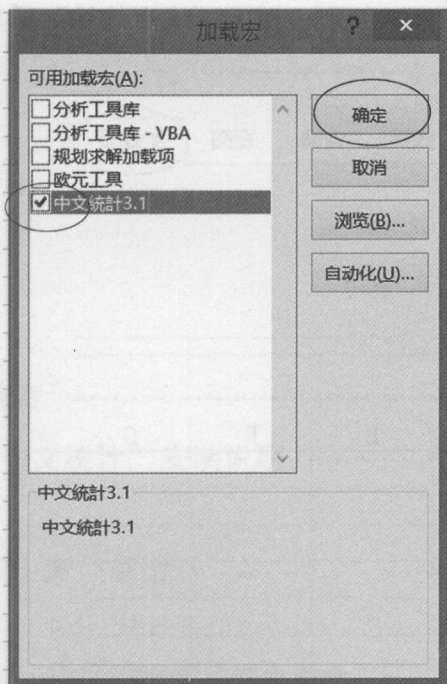


图 A-4 安装操作 4

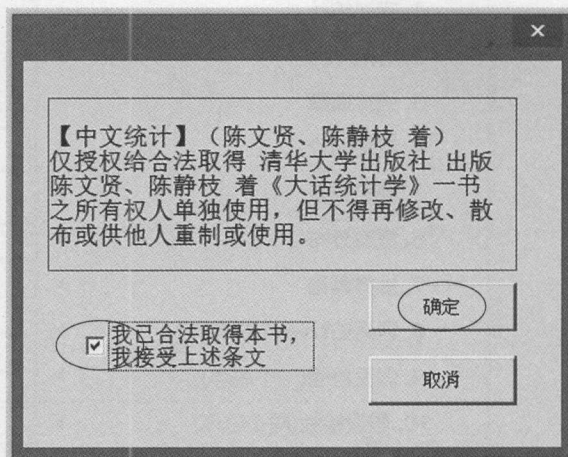


图 A-5 安装操作 5

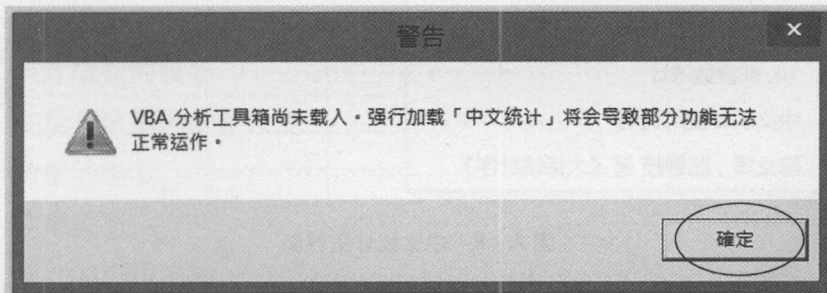


图 A-6 安装操作 6

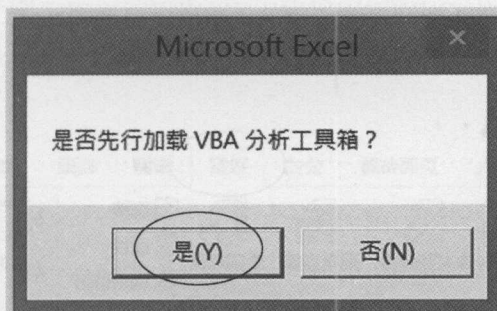


图 A-7 安装操作 7

(6) Excel 的菜单栏中出现“加载项”，单击“加载项”，就有“中文统计”菜单（选单），如图 A-8 所示。



图 A-8 中文统计主界面

如果菜单栏中没出现“加载项”，则关闭 Excel，再重新打开 Excel，就会出现“加载项”，此时可以开始使用“中文统计”。

中文统计需要用到 Excel 的“数据分析”，请检查是否已安装“数据分析”，如图 A-9 所示。



图 A-9 Excel 数据分析



前言	i
“中文统计” 安装说明	v
第1章 总论	1
1.1 统计是什么	3
1.2 统计学的基本概念	5
1.3 统计学的分类	8
1.4 抽样误差	10
1.5 统计数据的收集	12
1.6 变量与数据的衡量尺度	13
1.7 数据的类型	15
1.8 因果关系	16
1.9 统计的应用步骤	17
1.10 本章流程图	20
1.11 本章思维导图	21
习 题	22
第2章 描述统计	25
2.1 描述统计	27
2.2 统计表	27
2.3 统计图	29
2.4 总量指标与相对指标	38
2.5 平均指标集中趋势量数	39

2.6	相对位置量数	46
2.7	离差量数	50
2.8	形态量数	55
2.9	中文统计应用	59
2.10	本章流程图	66
2.11	本章思维导图	67
	习 题	68
第3章	时间序列	69
3.1	时间序列的分类	71
3.2	水平分析与速度分析	72
3.3	时间序列构成因素	74
3.4	平稳型序列预测	76
3.5	趋势型序列预测	79
3.6	季节指数分析	84
3.7	时间序列预测方法：趋势加季节	87
3.8	时间序列预测方法：趋势乘季节	88
3.9	预测误差	90
3.10	中文统计应用	92
3.11	本章流程图	98
3.12	本章思维导图	99
	习 题	100
第4章	统计指数	103
4.1	指数的意义与分类	105
4.2	总指数的编制	109
4.3	指数的性质	115
4.4	指数体系与因素分析	119
4.5	指数应用	123
4.6	中文统计应用	123
4.7	本章流程图	125

4.8 本章思维导图	126
习 题	127
第5章 概率理论	129
5.1 试验与样本空间	131
5.2 事件概率	133
5.3 排列组合的公式	136
5.4 事件概率的计算	140
5.5 条件概率	142
5.6 独立事件与互斥事件	146
5.7 贝叶斯公式	150
5.8 中文统计应用	153
5.9 本章流程图	155
5.10 本章思维导图	156
习 题	157
第6章 随机变量	159
6.1 随机变量	161
6.2 概率分布函数与概率密度函数	163
6.3 期望与方差	166
6.4 双随机变量	169
6.5 中文统计应用	177
6.6 本章流程图	179
6.7 本章思维导图	180
习 题	181
第7章 概率分布	183
7.1 离散型随机变量的概率分布	185
7.2 连续型随机变量的概率分布	196
7.3 正态分布概率的计算	205
7.4 中文统计应用	207

7.5 本章思维导图	209
7.6 本章流程图	210
习 题	211
第8章 抽样理论	213
8.1 随机抽样	215
8.2 统计量	216
8.3 抽样平均与抽样方差的概率分布	217
8.4 中心极限定理	221
8.5 分层抽样	223
8.6 整群抽样	224
8.7 系统抽样	225
8.8 中文统计应用	226
8.9 本章流程图	229
8.10 本章思维导图	230
习 题	231
第9章 统计估计	233
9.1 估计量	235
9.2 正态分布平均数与方差的点估计	237
9.3 总体平均数的区间估计	237
9.4 总体比例的区间估计	240
9.5 总体方差的区间估计	240
9.6 抽样的样本量	241
9.7 标准误差	243
9.8 中文统计应用	244
9.9 本章流程图	248
9.10 本章思维导图	249
习 题	250

第 10 章 统计检验	251
10.1 假设检验	253
10.2 计算第一类错误与第二类错误	256
10.3 假设检验的步骤与方法	261
10.4 假设检验的样本量	263
10.5 总体平均数检验, 方差已知	265
10.6 总体平均数检验, 方差未知	266
10.7 总体比例检验	267
10.8 总体方差检验	268
10.9 中文统计应用	268
10.10 本章流程图	272
10.11 本章思维导图	273
习 题	273
第 11 章 两总体估计检验	275
11.1 因果与相关	277
11.2 两个总体参数的区间估计	278
11.3 两个总体平均数检验, 方差已知	282
11.4 两个总体平均数检验, 方差未知但相等	283
11.5 两个总体平均数检验, 方差未知且不等	284
11.6 两个总体平均数检验, 样本是配对数据	285
11.7 两个总体方差检验	286
11.8 两个总体比例检验	287
11.9 中文统计应用	288
11.10 本章流程图	290
11.11 本章思维导图	291
习 题	292
第 12 章 方差分析	295
12.1 方差分析介绍	297
12.2 单因素方差分析, 样本量相等	299

12.3	单因素方差分析, 样本量不等	304
12.4	多重比较法	305
12.5	检验方差是否相等	306
12.6	参数估计	307
12.7	双因素方差分析, 无交互作用	308
12.8	中文统计应用	311
12.9	本章流程图	314
12.10	本章思维导图	315
习 题		316
第 13 章 回归与相关分析		317
13.1	回归与相关分析的区别	319
13.2	数学符号与关系式	321
13.3	一元线性回归分析参数的点估计	322
13.4	相关分析	324
13.5	检验自变量与因变量是否线性相关	328
13.6	回归与相关分析参数的区间估计与检验	329
13.7	中文统计应用	336
13.8	本章流程图	337
13.9	本章思维导图	338
习 题		339
第 14 章 分类数据分析		341
14.1	卡方检验	343
14.2	多项分布卡方检验	343
14.3	拟合优度检验, 分布的参数已知	345
14.4	拟合优度检验, 分布的参数未知	347
14.5	卡方检验独立性与同构性	350
14.6	中文统计应用	355
14.7	本章流程图	357
14.8	本章思维导图	358
习 题		358

第 15 章 非参数统计分析	361
15.1 非参数统计分析	363
15.2 符号检验	365
15.3 符号秩检验	369
15.4 游程检验, 检验随机性	371
15.5 Mann - Whitney 检验	373
15.6 Kruskal - Wallis 检验	375
15.7 Friedman 检验	376
15.8 Spearman 秩相关系数	377
15.9 中文统计应用	380
15.10 本章流程图	384
15.11 本章思维导图	388
习 题	389
第 16 章 结语	391
16.1 统计问题分类	393
16.2 误差名词说明	396
16.3 参数与统计量	398
16.4 统计概念复习	399
附录一 正态分布概率表	403
参考书目	405



第1章

总论

不明于计数，而欲举大事，犹无舟楫而欲经于水险也。

——管仲《管子》

如果你不能测量，你就不能管理。

——戴明 (W. E. Deming) 和德鲁克 (P. F. Drucker)

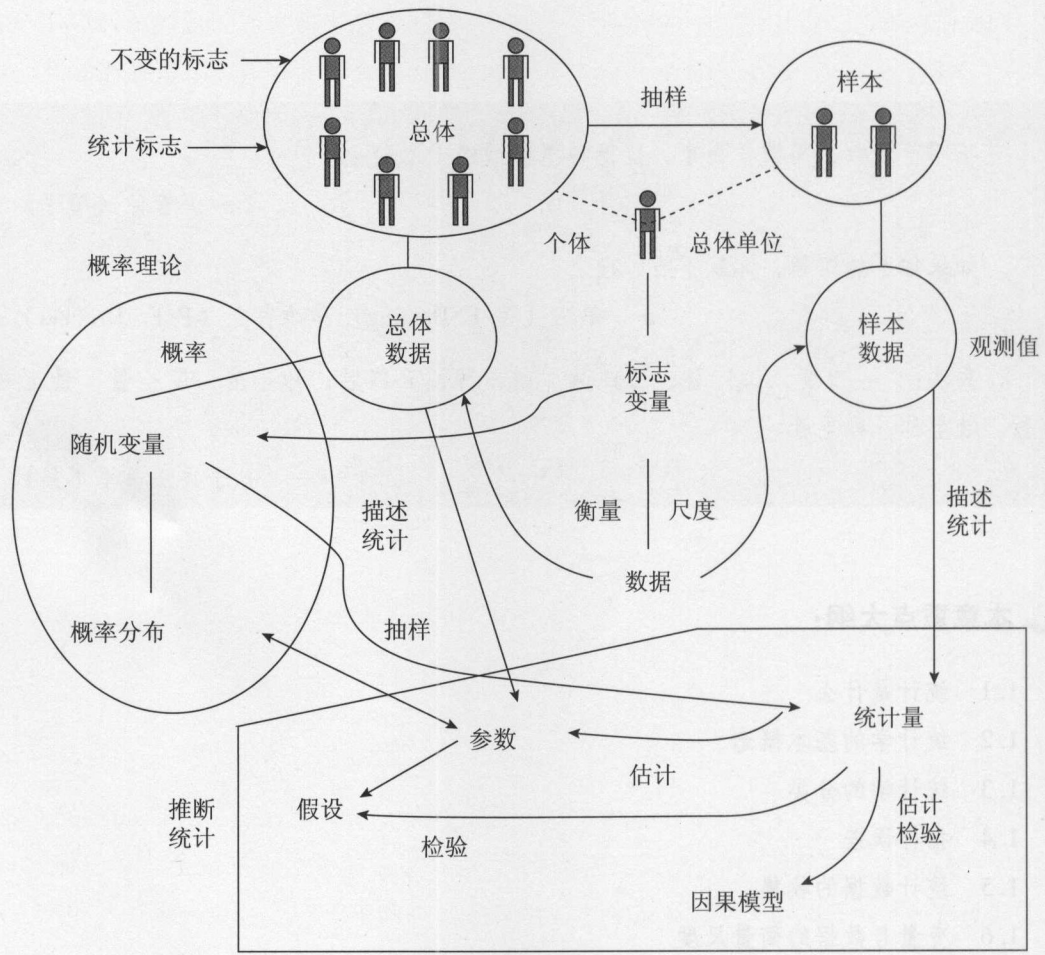
兵法：“一曰度，二曰量，三曰数，四曰称，五曰胜；地生度，度生量，量生数，数生称，称生胜。”

——《孙子兵法·军形篇》



本章重点大纲：

- 1.1 统计是什么
- 1.2 统计学的基本概念
- 1.3 统计学的分类
- 1.4 抽样误差
- 1.5 统计数据的收集
- 1.6 变量与数据的衡量尺度
- 1.7 数据的类型
- 1.8 因果关系
- 1.9 统计的应用步骤
- 1.10 本章流程图
- 1.11 本章思维导图



本章概念图

1.1 统计是什么

统计作为一门科学是从17世纪开始,起源于国情调查,有政治算数学派、国势学派以及古典概率学派,近代统计学演进到社会统计学派和数理统计学派。现代统计应用在各个领域:管理、工程、医学、农业、经济、社会、生物、气象、政治、军事等各学科。

统计一词,包括:统计工作、统计资料和统计学。本书的重点是统计学。

(1) 统计工作:统计的实践,应用统计问题,统计设计、搜集、整理、分析。

(2) 统计资料:统计工作获得的各种有关数据和信息,没有数据,就没有统计。

(3) 统计学:统计理论、分析数据、选择分析模型、了解计算结果、信息价值。

统计应用最多的是管理,但是文、理、工、商、医、农、政、经、社、法、军等各学科,都是广义的管理。

管理大师德鲁克和质量大师戴明说:“如果你不能测量,你就不能管理。”(If you can't measure it, you can't manage it)因为没有测量产生数据(data),就没有管理。有测量过的,才能管理。

上一句话反过来说:“你管理过的,一定要衡量。”管理过的衡量是评估绩效,例如:营销广告的绩效、信息系统的绩效、研究发展的绩效、投资的绩效,等等。

所以,管理要根据数据,然后要有绩效衡量:是否达到目的?是否产生价值?

测量(measure) → 数据(data) → 统计(statistics) → 管理 → 衡量(evaluate) → 绩效
(统计工作) (统计资料) (统计学) (统计工作与资料)

人们对于“统计”有好几种看法。有人认为统计是平均数、成数、表格和图形。有人认为统计是处理信息的程序和方法。有人说统计是以部分(样本)推论全体(总体)。还有人认为统计是对付与思考那些不规则发生的样本数据和不确定性的概率事件。也有人称统计是有关决策的科学。这些说法都对。

根据系统理论,一般系统有投入、处理、产出,以达到目标。从统计的投入(input:不规则发生的样本数据和不确定性的概率事件),处理(process:推论、程序和方法),产出(output:平均数、成数、图表),以及目标(objective:决策的科学)的观点来看统计。

数据或资料(data)是投入,信息(information)是产出。

统计学最主要的用途是:叙述已知(抽丝剥茧),与推论未知,以作为决策。统计学包括:①收集数据(定义变量、实验或调查);②表达数据(表格、图形);③将数据处

理为信息（百分比、平均数）；④思考概率问题；⑤产生结论预测或决策（以样本推论总体之估计、检验、及因果、关系）。

统计学的目的有以下4个。

（1）了解现象：描述统计是了解数据的呈现与性质，集中趋势的代表值或变异程度的离差值；时间序列和指数是了解变化因素和幅度。

（2）推测总体：统计检验和估计是推测总体。

（3）知道因果：两总体检验、方差分析、回归分析是知道因果。

（4）预测未来：时间序列是预测未来。

例题 1.1 是否新的可口可乐味道比较好？（了解现象，推测总体）

在1985年，可口可乐公司宣布，更改从1886年以来制造可乐的秘密配方。当新的可口可乐上市，《消费者报导》希望解答下列问题：是否新的可口可乐味道比较受欢迎？它跟对手百事可乐的比较又如何？《消费者报导》的研究人员找来95位同仁，分别尝试3种可乐，而杯子未注明品牌，让他们说出那一杯味道比较好。结果是：百事可乐和新可口可乐的偏好差不多；前两者比旧可口可乐多出一倍（2:1）的偏好。以上结果和新可口可乐刚上市的市场反应大相径庭，市场上反而不大能接受新可口可乐，旧可口可乐在很多地区的销售量还远大于新可口可乐。

例题 1.2 每周工作超时，中风风险大增？（知道因果）

2015年8月20日英国权威医学期刊《刺胳针》（The Lancet）刊登伦敦大学学院流行病学基维马吉教授根据17份研究调查，涵盖52万8908名男女，样本追踪时间平均长达7.2年，考虑吸烟、饮酒和身体活动程度等因素，研究工作时数与中风风险增加概率的关系。

研究发现，比起每周工作标准工时（每周工作35~40h）的人，那些每周工作41~48h的人中风风险高出10%，每周工作49~54h的人中风风险更是大增27%，而每周工作55h以上的人中风风险增加33%。研究也发现，即使考虑包括年龄、性别和地位在内等风险因子，长工时也使罹患冠状动脉心脏疾病的风险提高13%。

例题 1.3 二手烟是否对不吸烟者有害？（知道因果）

20世纪80年代美国加州大学圣地亚哥分校做了一个肺功能试验。这个试验是检查200位不吸烟的中年人，他们通常处于不吸烟的环境；另外200位也不吸烟的中年人，但是20年来，他们经常在吸烟的环境中工作。这两组人同时和吸烟者的肺功能比较。检验结果是：吸二手烟和不吸二手烟的不吸烟者，在肺活量与吐气率两方面，无显著差异。但是在肺部的细部呼吸方面，吸二手烟者比不吸二手烟者，有显著的损害。这个研究建议：

长期暴露在二手烟的环境中，对健康是有害的，会显著降低肺部的细部呼吸的功能。

例题 1.4 民意调查可靠吗？（推测总体、预测未来）

1936 年美国总统大选，共和党的蓝登（Landon）对上民主党的罗斯福（Roosevelt）。《文学文摘》（Literary Digest）根据其读者名册、电话号码簿（当时只有 1/4 家庭有电话）、汽车注册名单、杂志读者名册和俱乐部会员名单，寄出 1000 万份问卷，回收 230 万份，预测蓝登以 57 : 43 胜罗斯福。同时间，一个叫盖洛普（George Gallup）的年轻人只抽样了 5 万人，预测罗斯福会赢，被嘲笑太天真，因为样本量太少。结果，罗斯福赢了，得票率 62%！（《文学文摘》不久因此破产了，盖洛普则成为专业的民调专家）时至今日，民意调查的样本量，只要一两千个，就可以推估千万人的总体参数，数万样本可以增加的准确度很少。

1948 年，美国总统大选，共和党杜威（Dewey）与民主党杜鲁门（Truman）竞选总统。选举前，盖洛普民意调查（Gallup Poll）显示，共和党的杜威领先。1948 年 11 月 3 日，《芝加哥论坛报》刊出一个惊人的标题“杜威击败杜鲁门”。杜鲁门反而拿该报纸造势，争取同情票，结果扭转乾坤，杜鲁门赢得选举，当选总统。但是在 1948 年 11 月 3 日，是否真的是杜威领先杜鲁门，也无从查证。（这就是未知的参数）

1.2 统计学的基本概念

定义 总体（population）是要研究的数据的全体对象。

我们要研究公司的薪资所得，则全体员工就是总体。

定义 对“全部”总体进行调查，称为总体普查（population census）。

通常总体普查要花费相当大的人力、时间与金钱。有时要找到全部总体非常困难。对于质量管理的检验，有的是破坏性检验，总体普查以后，全部产品都报销。

定义 总体的基本成份，称为个体或总体单位（unit）。

个体或单位可能是人、动物或商店等。例如：学生、产品、员工、消费者等。

定义 取出总体的“部分”个体，称为抽样（sampling）。抽样出来的个体集合，称为样本（sample）。

定义 样本的数目称为样本量或样本容量（sample size）。

根据抽样的方法，总体分为有限总体和无限总体，如果总体单位的数目是有限的，每个样本抽出后不放回（不重复），且样本量占总体单位数目的比例大于10%，则为有限总体。

定义 标志是总体单位的属性和特征（characteristic）的名称。

标志有不变标志和可变标志。不变标志是总体构成的基础，例如：两个总体检验，分辨两个总体的标志，如性别、地区、处理方法等，是不变标志。可变标志是要进行统计（叙述、概率、推论）的个体的特征。例如：学生的成绩、产品的质量、员工的薪资、消费者购买的品牌、零件是否为良品、选民支持的候选人等。

标志又分为质量标志和数量标志，质量标志是定性的标志，数量标志是定量的标志。上述品牌、良品、支持的候选人是质量标志。成绩、质量、薪资是数量标志。图1-1中的方差分析检验不同教师的学生成绩的平均数是否相等。教师称为“因素”就是质量标志。不同的教师是因素的“水平”，可视为不同的总体，每个总体的教师是不变标志。“观测值”的名称（学生成绩）就是数量标志。

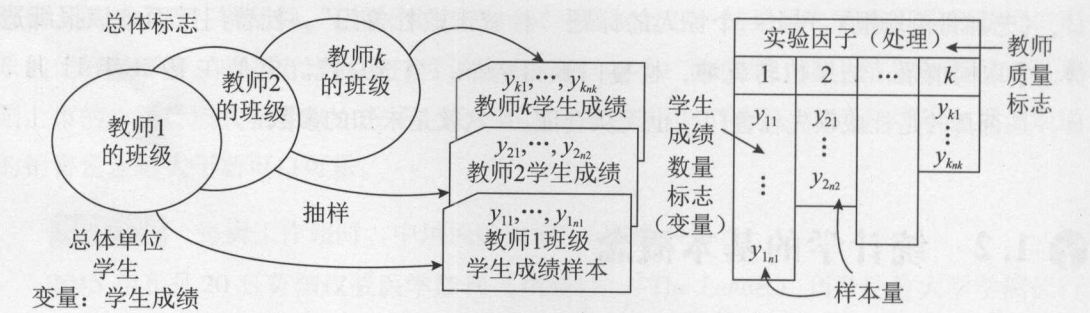


图 1-1 质量标志和数量标志

标志有标志名称，如教师、学生成绩；有标志值，如教师1、教师2、学生成绩值。

定义 统计指标（indicator）是说明总体数量特征的名称及数值，例如国内生产总值、总人口数等。

在第2章描述统计，量数（measure）是总体的数量特征也是指标，例如集中趋势量数。

指标又分为：总量指标、相对指标、平均指标和变异指标。完整的指标应具备：时间限制、空间限制、指标名称、统计数值、计量单位等5个构成要素。（下一页说明）

定义 变量（variable）是可变标志和统计指标，前者是总体单位（个体）的变量，后者是总体的变量。

本书的变量，多数是指个体的变量。例如：总体是某一班级（不变标志）的学生，变量是学生的性别、分数、身高等。

数不知道。中国台湾报考托业的高中生不能代表所有高中生总体，中国大陆报考托业的考生是否代表所有中国大陆大学生的总体？

例题 1.5 说明完整的指标应具备：时间限制（2012 年）、空间限制（中国台湾的大学生）、指标名称（平均总成绩）、统计数值（504）、计量单位（分数）等 5 个构成要素。其实还有一个构成要素就是容量限制：总体容量和样本容量（每年大学生的人数和报考人数）。

常用的参数和统计量的符号如表 1-1 所示，更完整的参数和统计量的符号，如表 16-2 所示。

表 1-1 总体与样本的符号

指标名称	总体 参数符号	样本 统计量符号	计量单位
容量（数目）	N	n	无单位：整数
（算术）平均数	μ	\bar{x}	定距尺度： x 的单位
几何平均数	G	G	定比尺度的百分比
调和平均数	H	H	相对单位：速度，单价
比例	π	p	无单位：百分比
方差	σ^2	s^2	同 x 的单位的平方
标准差	σ	s	同 x 的单位
中位数	M_ρ	M_ρ	同 x 的单位
协方差	$Cov(X, Y) = \sigma_{XY}$	$Cov(x, y) = q_{XY}$	x 单位 \times y 单位
相关系数	ρ	r	无单位
指数	P, Q	P, Q	无单位

1.3 统计学的分类

统计学的内容，分成两大类：描述统计（descriptive statistics）与推断统计（inference statistics）。描述统计是探讨总体数据的性质或样本数据的性质，是将数据加以组织分析，并且用图形或数值（指标）表达一些现象，描述某些关心的主题，例如：集中代表值，离散程度，分布型态。指数和时间序列归类在描述统计，主要是将历史数据整理成一个信息（变动比率、趋势值、季节值等指标）。至于时间序列的回归预测，则应该属于推断统计的范围。

探讨总体与样本之间性质的另一方向是推断统计：利用样本数据，加上推论或归纳，得到总体未知参数的估计或检验。推断统计是利用概率分布的理论以及估计、检验、预测等方法，利用抽样的有限数据，来归纳或推断总体的一般性质。在工程、医学、管理等领

域，或在日常生活，我们都无法掌握完全的信息，来知道事实的全部真相。我们都是利用有限和不完全的信息在做决策与推论。所以推断统计是，以有限的信息，来了解与处理，周遭的不确定事件，或者是已经存在但未知的事实（总体参数）。

在描述统计和推断统计之间，扮演串场角色的是概率理论。概率理论是：总体参数已知，利用演绎或仿真，得出样本或事件（总体的部分集合）的概率，再利用随机变量，定义概率分布函数，于是得到抽样统计量的概率分布，然后再将其应用到推断统计。

- 统计学分类的关系说明如下（第 8，9 章的统计量对应第 2 章的指标公式）：
- 总体→标志、变量→数据、尺度→指标、参数、表格、图形（第 2 ~4 章描述统计）
 - 总体 + 已知参数→概率、随机变量、概率分布、抽样→统计量（第 5 ~8 章概率理论）
 - 样本数据→统计量→估计、检验→总体参数或因果关系（第 9 ~15 章推断统计）

另外，统计学又可分为理论统计和应用统计。理论统计是概率理论和数理统计，数理统计学则是讨论应用统计背后的理论基础的学科。应用统计分为描述统计和推断统计，是应用在各学科的统计学，例如商业统计、生物统计、农业统计、社会统计等。

图 1-3 是统计学的内容分类和中文统计菜单（选单）。

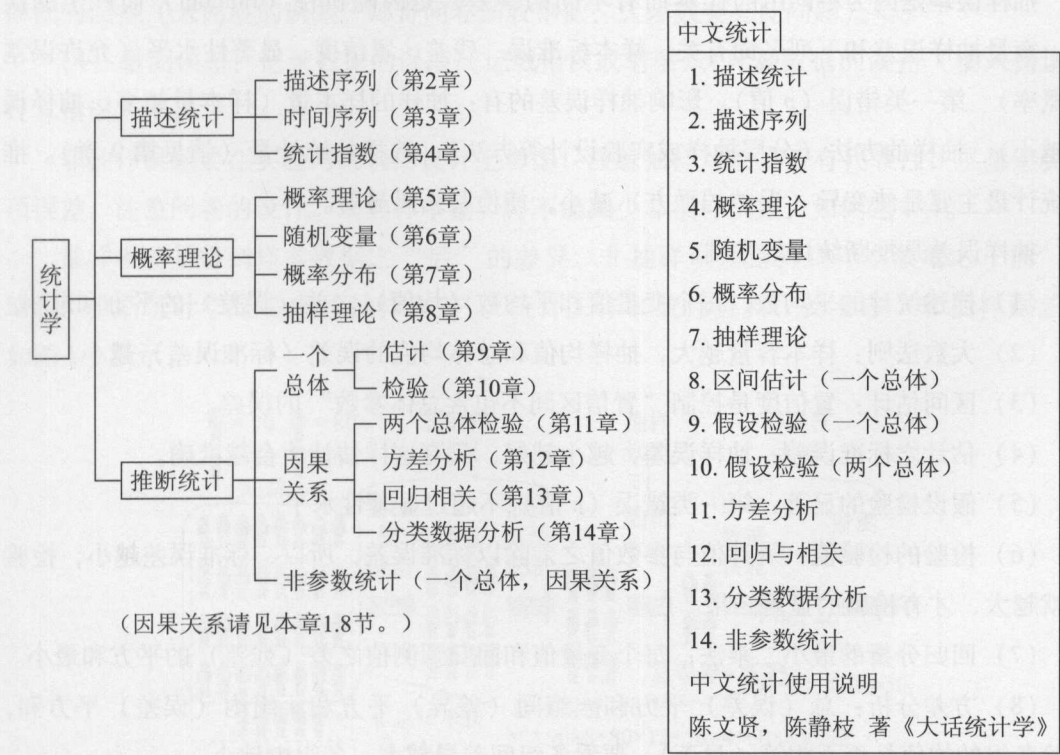


图 1-3 统计学的内容分类和中文统计菜单（选单）

1.4 抽样误差

推断统计主要建立在3个基础之上：①总体参数与样本统计量；②抽样随机性（概率理论）；③抽样误差。总体参数是推断统计的目的（What，要做什么），概率理论是原因（Why，为什么），抽样误差是方法（How，如何做）。这是5W1H（What，Why，Who，When，Where，How）中，最重要的3个：Know What，Know Why，Know How。

因为抽样数据只是总体数据的一部分，所以抽样数据计算出来的统计值（例如：样本平均数），与总体数据计算出来的参数值（例如：总体平均），会不相等，其差异是误差。

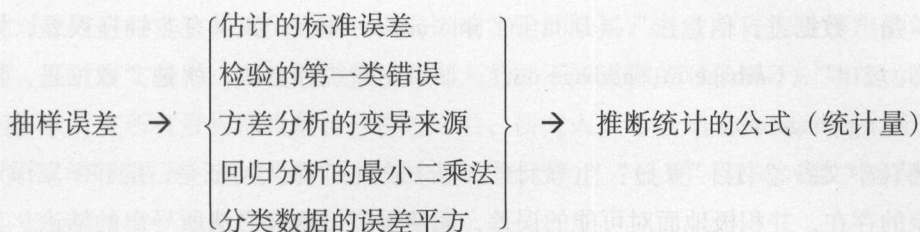
误差是“样本数据统计值与总体的参数值”的差（绝对值），误差分为抽样误差和非抽样误差。

定义 抽样误差（sampling error）是因为抽样的“随机性”所造成的误差。

抽样误差是因为不同组的样本而有不同的误差，是因随机性（random）而产生的误差。衡量抽样误差和下列名词有关：样本标准误、残差、置信度、显著性水平（允许误差的概率）、第一类错误（ p 值）。影响抽样误差的有：抽样的样本量（样本量越多，抽样误差越小）、抽样的方法（分层抽样或实验设计等方法）、选择的统计量（请见第9章）。推断统计最主要是使变异（误差的平方）减小，使检验结果显著。

抽样误差是推断统计的基础。

- (1) 描述统计的平均数：每个变量值和平均数（均值）之差（误差）的平方和最小。
- (2) 大数法则：样本容量越大，抽样均值和总体均值的误差（标准误差）越小。
- (3) 区间估计：置信度是控制“置信区间不包含总体参数”的误差。
- (4) 估计之标准误差：抽样误差，越小越好，只有这样估计才会越准确。
- (5) 假设检验的显著：第一类错误（ p 值）不超过显著性水平。
- (6) 检验的检验值：统计值与参数值之差除以标准误差，所以，标准误差越小，检验值就越大，才有检验的显著结果。
- (7) 回归分析的最小二乘法：每个变量值和回归预测值之差（残差）的平方和最小。
- (8) 方差分析：总（误差）平方和 = 组间（差异）平方和 + 组内（误差）平方和，检验各组的均值是否不相等（显著），要看各组间差异越大，各组内越小。
- (9) 分类数据分析：以样本值和理论值之差（误差）的平方和，检验一个变量的概率，或两个变量的独立性。



定义 非抽样误差 (non-sampling error) 是在抽样过程中, 由于人为错误而造成的误差。

非抽样误差是因“人”(研究者或受测者)而产生的误差。非抽样误差包括以下几种。

(1) 选择样本抽样框 (sampling frame) 的错误, 样本不能代表总体。抽样框是抽样个体的名册, 用来抽选样本的个体, 如: 电话簿名册、毕业纪念册、会员名单等。

(2) 选择抽样方法的误差, 选择抽样的方法包括: 便利抽样, 以最方便的方法选择样本, 如街头调查、利用学生作实验; 自发性响应样本, 样本以自动应答的方式取得, 如电视台的叩应 (call-in) 或报纸、杂志、博客、BBS 的来应 (write-in), 其回答的样本都是有心人; 还有例题 1.5 报考托业的中国台湾高中生, 以上都不能代表总体。

(3) 取得数据的误差: 问卷设计得不好, 问题敏感, 受访者不愿答或故意答错, 回收率低的误差 (未回应的误差, 邮寄问卷回收率低, 大多数会有此问题)。

(4) 量测误差: 记录数据的误差 (记载错误或笔误)、计算数据的误差 (输入错误或计算错误) 等。

非抽样误差要在实验与调查的设计上考虑, 注意抽样对象是否有代表性, 尽量避免这项误差。注意问卷的设计。增加样本量, 并不能减少非抽样误差, 如例题 1.4。

抽样误差是得到样本数据之“后”的差异。非抽样误差是得到样本数据之“前”的错误。推断统计学是考虑“抽样误差”。统计工作和统计资料, 要考虑“非抽样误差”, 如图 1-4 所示。

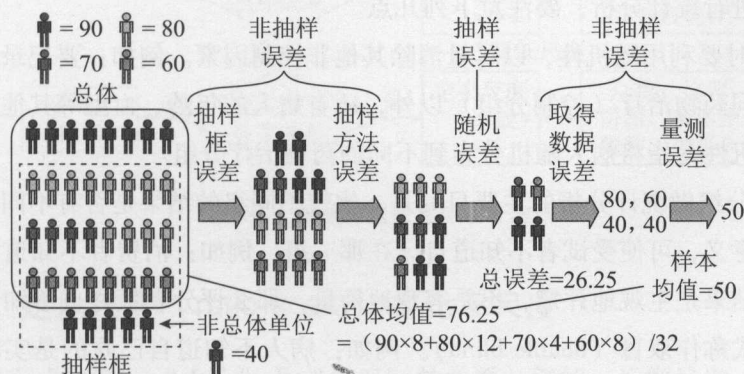


图 1-4 抽样误差与非抽样误差

统计学是“数据进，信息出”（Data in, information out）；如果有非抽样误差，则是“垃圾进，垃圾出”（Garbage in, garbage out）；如果用错统计方法，就是“数据进，垃圾出”（Data in, garbage out）。

赵民德（中文参考书目 [7]）：“（统计学）方法的一个最大特征是：统计学家深切地体会到误差的存在，并积极地面面对可能的误差，而使得经过这套方法所导出的结论，其因误差而产生的暧昧减少。统计学的方法并不能无中生有，但它的确致力于尽量滤去误差，而得到传统方法所不能得到的结论。误差如水，真相若石。水落，所以石出。如果水中原本无石，水落当然也仍然无石。统计的方法之所以大行其道，是因为误差是近代生活的一部分。一切人类所搜集的数量数据中，其不包含误差者百不得一。社会越进步，则所需要搜集与分析的（包含着误差的）数据也越多。而我们越尝试地去以无误差的概念去推演结论，所得的结论里便越充满误差。而统计方法，因为能正视误差的存在，反而可以得到更合理的结论。……统计方法：围绕着包含了误差的数字，所作的种种精巧的努力。”

以上所说的“误差”是“抽样误差”。本书第 16 章结语，介绍了“误差”的名词与关联性。

1.5 统计数据的收集

要得到样本数据，可以通过实验与调查（包括观察）两种方式。

定义 实验（experiment）是对样本，加以控制（control）分组，再进行测量或观察。

例如：医学实验，将病人分成两组或两组以上，利用不同的药物治疗（控制），再观察其病情。实验数据通常经由方差分析或双总体统计检验，做统计分析。

利用实验进行统计分析，要注意下列几点。

（1）分组时要利用随机性，以尽量消除其他非控制因素。例如：要记录病人吃药后的病情，除了不同药物治疗（控制分组）以外，还有病人的年龄、血型等其他因素，会影响病情。所以随机性是指将病人随机分布到不同的药物治疗分组。

（2）实验分组做统计分析的主要目的是，比较不同组的结果是否有不同。为了使其结果比较客观有意义，可使受试者不知道自己在哪一组，例如：消费者不知道自己用的是哪一品牌。如果结果是主观地评定，不是客观地衡量，那么评分者最好也不知道受试者是哪一组，这种方式称作双盲（double blind）。例如：病人不知道自己吃的是实验药或维他命片；同时，医生或护士也不知道病人是哪一组。总之，双重隐瞒是消除实验中可能的个人感情因素，以避免影响实验结果。双盲实验通常在实验对象为人类时使用，目的是避免实

验的对象或进行实验的人员的主观偏向影响实验的结果，通常双盲实验得出的结果会更为严谨。在双盲实验中，实验的对象及研究人员并不知道哪些对象属于对照组，哪些属于实验组。只有在所有资料都收集及分析过之后，研究人员才会知道实验对象所属组别，即为“解盲”。解盲结果，若主要疗效指标未呈现统计学上“显著”意义，则“解盲失败”。

实验要注意：随机性分组与双重隐瞒。

定义 调查（survey）是对总体或样本，不加以控制（control）分组，直接进行访问或观察。

例如：市场问卷调查、电话访问、座谈或个人访问等都是调查。调查方式有：自我测验、访问、电话等，必要时先做试测。

1.6 变量与数据的衡量尺度

数据衡量是将变量给予一个实数值（观测值），但是因为变量的性质不同，所以有不同的衡量尺度。以下我们介绍4种衡量尺度，如图1-5所示。

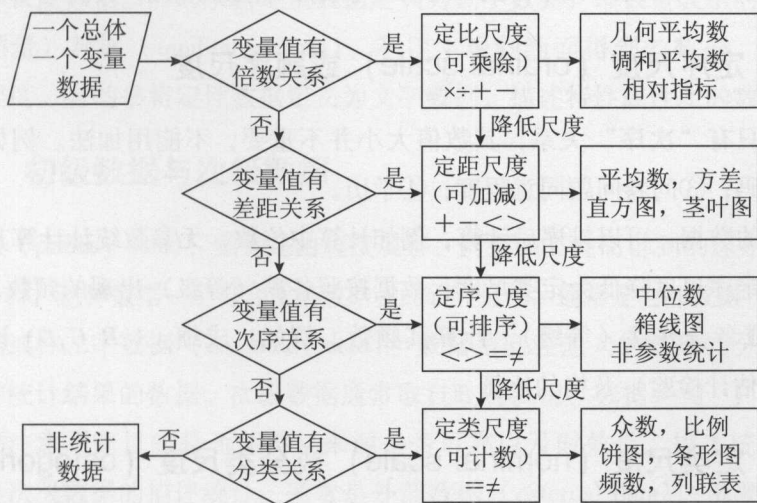


图 1-5 数据衡量尺度与统计方法

1.6.1 定比尺度（ratio scale）或比率尺度

数据之间有“次序”、“大小”及“比率”的关系。例如：长度尺度、质量尺度、绝对温度、金钱、面积、体积、时间间隔等。同时可以定义一个原点，即零值“0”，零值表

示“没有”，没有长度、质量、温度、钱等。

定比尺度的数据做分析时，可以经过数学运算（+、-、 \times 、 \div ）及转换（ X_i^2 或 $\log X_i$ ）等。

定比尺度的数据常用的代表值是平均数。在统计推论中，是估计及检验其算术平均数。用定比尺度的数据制作统计图时，有时会故意改变尺度比率，或故意忽略原点，或用二度空间尺度面积表示，但可能造成误解。

1.6.2 定距尺度 (interval scale) 或区间尺度

数据之间有“次序”及“大小”的关系，而没有“比率”的关系。主要是衡量一个数据大于另一个数据多少，而并非其倍数关系。

定距尺度的原点“0”并非代表“无”（无温度，无智商，无尺寸）。例如：①温度尺度（ $^{\circ}\text{C}$ 或 $^{\circ}\text{F}$ ）；②智商分数；③衣服或鞋子的号码。

定距尺度的数据有基本的测量单位，可以计算其数值的加减。定距尺度的数据常用的代表值是平均数。在统计推论中，也是估计及检验其算术平均数。

定距尺度的数据也可以用定序尺度（非参数）的统计方法。定距尺度降低为定序尺度： n 个数据顺序排列，从小到大，给予 1 到 n 的秩（等级），若数值相同，其秩（等级）作平均。

1.6.3 定序尺度 (ordinal scale) 或顺序尺度

数据之间只有“次序”关系，其数值大小并不重要，不能用加法。例如：①考绩等级；②门牌号码；③问卷问题同意程度；④学历。

定序尺度的数据，可以排序后计算，例如计算中位数，无参数统计计算其秩（等级），检验中位数。定序尺度降低为定类尺度：数据按照各秩（等级）出现的频数，不管其顺序秩（等级）。或者合并秩（等级），计算其频数，例如：成绩（A, B, C, D）计算及格不及格，计算（或估计检验）及格的比例。

1.6.4 定类尺度 (nominal scale) 或分类尺度 (categorical scale)

数据之间没有任何“次序”，“大小”及“比率”的关系，只有“分类”关系。例如：①性别数据；②颜色数据；③电话号码、邮政编码、球员编号、职业别、地区别。

定类尺度的数据常用的代表值是众数。在统计推论中，定类尺度变量是计算总体比例，或在回归分析中当作虚拟变量（dummy variable）。

定比尺度和定距尺度的主要参数是平均数；定序尺度和定类尺度的主要参数是比例，如图 1-5 所示。

1.7 数据的类型

数据以不同的性质分类，有下列分类方法。

1.7.1 连续数据与离散数据

连续数据 (continuous data)：数据是可以有小数或分数。定比尺度和定距尺度通常是连续数据。计量值数据相当于连续数据。

离散数据 (discrete data)：数据是整数，没有小数或分数。定序尺度和定类尺度通常是离散数据；定比尺度也可能是离散数据，例如：生产的个数。计数值数据相当于离散数据。

1.7.2 定量数据与定性数据

定量（数量）数据 (quantitative data)：利用客观标准衡量而得到的数据。例如：产品寿命数据，长度数据。有的书将定量数据定义为数字数据，以数量表示的数据。

定性（质量）数据 (qualitative data)：利用主观判断而得到的数据。例如：考试数据，同意的程度。有的书将定性数据定义为文字数据，描述特性或性质的数据。

1.7.3 初级数据与次级数据

初级数据 (primary data)：数据是由直接观察、调查或实验而得到的原始数据，未经他人的整理或分析，这种数据一定符合搜集数据者的研究目的。通常是内部数据 (internal data)。

次级数据或称二手数据 (secondary data)：数据是已经他人的整理或分析，变成频数分布表或某种统计结果的数据。次级数据通常取自政府机构、数据公司、广告公司等。引用次级数据要注意研究目的是否相符、来源是否可靠以及时效性。描述统计中的分组数据，可以说是次级数据的描述统计。通常是外部数据 (external data)。例题 1.2 可以说是次级数据的研究。

1.7.4 横断数据与纵向数据

数据来源根据是否与时间相关，通常可分成横断（面）数据 (cross-sectional data) 及纵向数据或追踪调查数据 (longitudinal data or panel data)。横断数据是静态数据，收集一个时间点的数据，在“同一时间”的单总体、多总体或多变量的数据。纵向数据是动态

数据，是经过一段时间，收集“不同时间点”的数据，指数的数据和时间序列数据是纵向数据。只做一次式的调查，是横断数据；实验虽然要经过一段时间，但是如果只记录最后结果的数据，那么也是横断数据。

1.7.5 数据集合

记录或个案是个体单位的变量集合，记录和变量可以用一个电子表格（worksheet 或 spreadsheet）来显示，如 Excel。所以，本书所用的中文统计是建立在 Excel 上的一个加载项。数据电子表格相当于一个矩阵，行（row）代表记录，列（column）代表变量。

1.8 因果关系

在统计学中，可以利用两组数据（“两个变量”或“两个总体”），分析其因果关系或相关性。两总体的平均数或比例检验，其“因”是两总体的分类变量，例如“性别”或“地区”；其“果”是平均数或比例的变量，例如“成绩”或“候选人得票率”。所以，因果关系的假设检验是：同年龄的“男生和女生”的智商或成绩平均数，是否相等或有显著差异，即“性别”是否影响“智商或成绩”；地区是否影响候选人得票率；吸烟影响健康是确定的因果关系；出生月份决定未来的职业、健康与命运，你相信这个推断吗？

分类数据分析的卡方检验，其“因”是两类以上的定类变量，其“果”也是定类变量，例如：如表 1-2 所示，1912 年，泰坦尼克号撞上冰山而沉没，乘客和组员共 2223 人，死亡 1517 人，其中不同“性别”（因）的死亡率（果）是否有显著差异？不同“身份（旅客等级或组员）”（因）的“死亡率”（果），是否有显著差异？第 5 章的条件概率与第 14 章的分类数据分析将对此给予回答。

表 1-2 泰坦尼克号生死录

因 \ 果	头等舱		二等舱		三等舱		组员		总和	
	男	女	男	女	男	女	男	女	男	女
存活	54	145	15	104	69	105	194	20	332	374
	199		119		174		214		706	
死亡	119	11	142	24	417	119	682	3	1360	157
	130		166		536		685		1517	
总和	173	156	167	128	486	224	876	23	1692	531
	329		285		710		899		2223	

通常，第11~14章的原假设是“没有因果关系”，检验结果“拒绝原假设”表示有“显著差异”，所以“有显著差异”表示“有因果关系”。

回归分析就是两个变量的因果关系，检验自变量X（因）对因变量Y（果）的直线关系是否显著。例如：广告预算对销售额的影响是否显著，信息科技的支出对企业的获利绩效的影响是否显著。

不同数据尺度检验因果关系的统计方法也有不同。表1-3是从第11章开始到第15章，不同尺度的因果或相关的统计方法。

表 1-3 不同尺度的因果关系的统计方法

因 果	定类尺度		定序尺度	定距尺度
	2 分类	≥2 分类		
定类尺度	两个总体比例检验 游程检验 (run test)	列联表 分类数据分析		判别分析 Logistic 回归
定序尺度	符号检验 (sign test) Wilcoxon 符号秩检验 Wilcoxn 秩和检验	KW 检验 Friedman 检验	Spearman 检验	Spearman 检验
定距尺度	两个总体平均数检验 两个总体变异数检验	方差分析	时间序列指数	散点图、回归分析、 相关系数

1.9 统计的应用步骤

《孙子兵法·军形篇》中写道：“兵法：一曰度，二曰量，三曰数，四曰称，五曰胜；地生度，度生量，量生数，数生称，称生胜。”地：分析地形险易情况。度：判断战区战线区域。量：计划部署战场容量。数：决定所需人力数量。称：权衡比较双方优劣。胜：未战已经胜券在握。统计的应用步骤（如图1-6所示），和兵法不谋而合。

统计工作：

(1)（地）了解问题，定义总体、变量。总体是什么？有几个总体？（分类总体的标志是什么？）有什么变量？（要衡量总体的什么性质？）有几个变量？是否有两个以上变量的相关或因果关系？

(2)（度）认定变量值的数据尺度，决定指标、参数。数据的尺度是什么？什么指标？什么参数？描述统计或推断统计？

统计资料：

(3)（量）决定实验、调查、观察或二手数据。实验是抽样，调查决定普查或抽样。

设计实验步骤或调查方式。选择抽样方法，决定样本容量。

(4) (数) 收集数据，决定数据特性（符合假定条件如正态）、统计量、统计模型。辨认 (identify) 统计模型，检查假定条件，统计模型的假定条件有：数据尺度、正态分配、抽样独立性、方差条件等。

统计学：

(5) (称) 数据分析，普查是描述统计，选择表达的方式。抽样是推断统计，选择统计分析模型。表达方式有：表格、图形或代表值等。计算 (compute) 结果。

(6) (胜) 得到信息、报告结论，或导出决策。

解释 (interpret) 结果，得到信息、报告结论、实施决策、衡量决策的结果。

一般统计学教科书的例题或习题的解答步骤，通常已有数据，只要做第 4, 5, 6 步骤。

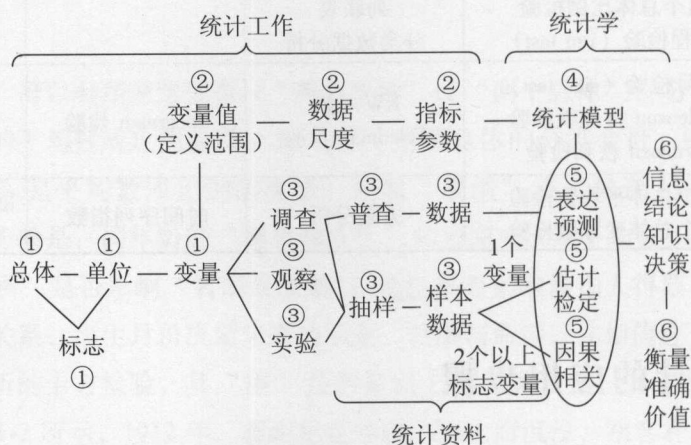


图 1-6 统计的应用步骤 (数字代表上述步骤)

例题 1.6 可乐独卖权利的决策 (参考英文书目 [14] Keller 2009)

T 大学有 30 000 学生，要和 P 可乐饮料公司签定独卖合约，在校园内只能卖 P 可乐，学校每年收 100 万元回馈金，加上 P 可乐全年销售金额的 30%。目前在 T 大学有 3 种以上的可乐销售，P 可乐每周平均销售 20 000 罐，但是不知道其他可乐的销售量。一年 40 周 (寒暑假没学生)，所以 P 可乐在 T 大学的年销售量为 800 000 罐。如果 P 可乐每罐售价 ¥3，每罐成本 ¥1。P 可乐公司有 2 周的考虑期间，请问该公司如何做决定？

解答：目前 (没有独卖) P 可乐每年获利 $20000 \times (\text{¥}3 - \text{¥}1) \times 40 = \text{¥}1\,600\,000$

假设 π = P 可乐在 T 大学的市场份额 (市占率)

则每年 T 大学可乐独卖的销售数量为 $X = 800\,000 \div \pi$ (罐)

P 可乐独卖每年的获利 $X \times \text{¥}3 \times 0.7 - X \times \text{¥}1 - \text{¥}1\,000\,000 = \text{¥}1.1X - \text{¥}1\,000\,000$

独卖优势 $A = \text{独卖每年的获利} - \text{没有独卖每年的获利} = \text{¥}1.1X - \text{¥}2\,600\,000$

$$A = 1.1 \times (800\,000 \div \pi) - 2\,600\,000 = 880\,000/\pi - 2\,600\,000$$

P 可乐在 T 大学的市场份额与独卖优势如表 1-4 所示，市场份额越高，独卖优势越少。

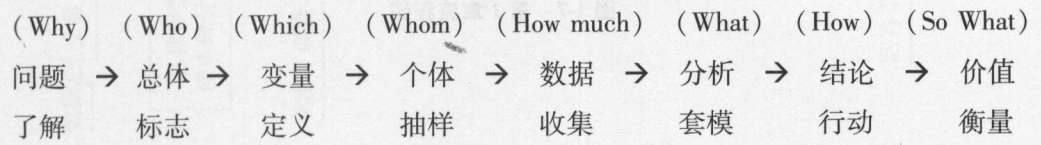
表 1-4 可乐在 T 大学的市场份额与独卖优势

市场份额 π	0.2	0.25	0.3	0.35	0.40	0.45	0.5
独卖优势 A (万)	180	92	33	-8.6	-40	-64	-84

当 P 可乐的市场份额为 33.85%，独卖优势为 0，为独卖的损益两平点。P 可乐的市场份额越低，和 T 大学签定独卖合约越有利。因此，P 可乐决定用一周的时间，进行统计推断，推断市场份额。统计学的应用步骤如下。

- (1) 总体 = 30000 学生，变量 = 每个学生每周购买可乐的品牌及数量，参数 = T 大学每年可乐的销售数量或 P 可乐的在 T 大学的市场份额。
- (2) 变量有定类尺度（可乐品牌）及计数值数据（可乐数量）。
- (3) 决定调查，抽样 500 个学生，记录每人一周买可乐的品牌及数量。
- (4) 每个样本数据如：{CCP} {PP} {CCCT} {PC} {T} 等，P 表示 P 可乐、C 表示 C 可乐、T 表示 T 可乐。计算 500 个学生中，P、C、R 的个别总和，及全部可乐的总和。
- (5) “P 的总和”除以“全部总和”即为 P 可乐的市场份额的估计值，这是点估计，根据这个市场份额的点估计，就可以决定是否和 T 大学签订独卖合约。如果要检验下列假设，则可能要多抽样几次 500 个学生。
原假设 $H_0: \pi \geq 33.85\%$ ， 不同意独卖合约。
备择假设 $H_1: \pi < 33.85\%$ ， 同意独卖合约。
- (6) 得到决策：是否同意独卖合约。
- (7) 进一步考虑因素：这个推论的假定条件是 T 大学全部可乐的销售量等于 P 可乐独卖的销售量。实际上其他牌可乐的忠诚度，使得独卖不见得将所有市场份额，都转为 P 可乐。例如 C 可乐爱好者，在独卖后，可能不会买 P 可乐。因此，“独卖后”P 可乐的销售数量，不等于 P 可乐销售数量除以“没有独卖的市场份额”。应该将每年独卖后的销售数量打折。
- (8) 问题：如果能够推导 π 的估计量或统计量的概率分布或方差（标准差），那么才可以进行第 5 步的假设检验推论。决策法则：如果 π 的点估计值大于 33.85%，则不同意独卖合约。

最后，将统计的应用步骤，再整理如下：（套模是套用模型）



1.10 本章流程图

本章流程图如图 1-7 所示。

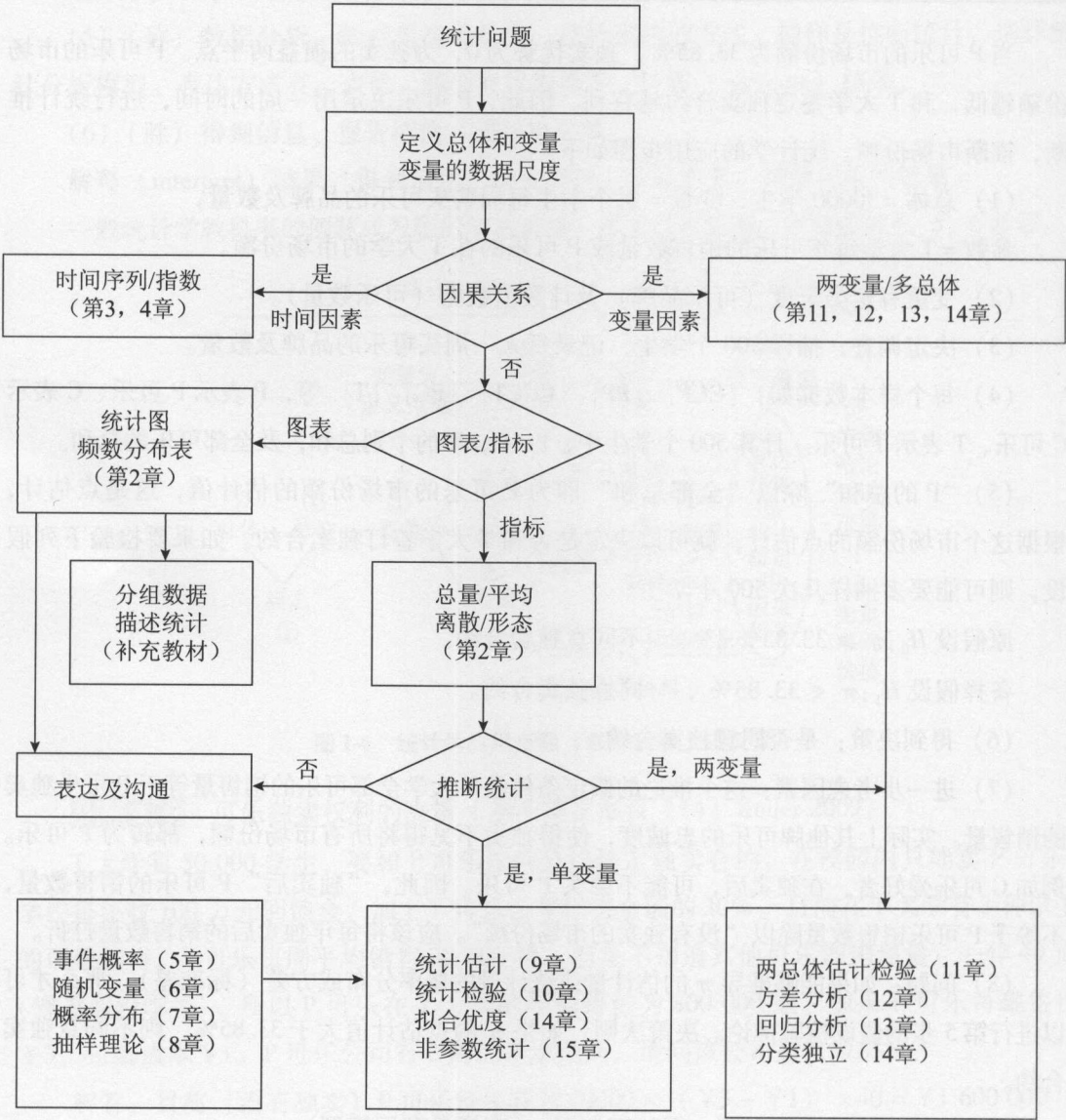


图 1-7 第 1 章流程图

1.11 本章思维导图

本章思维导图如图 1-8 所示。

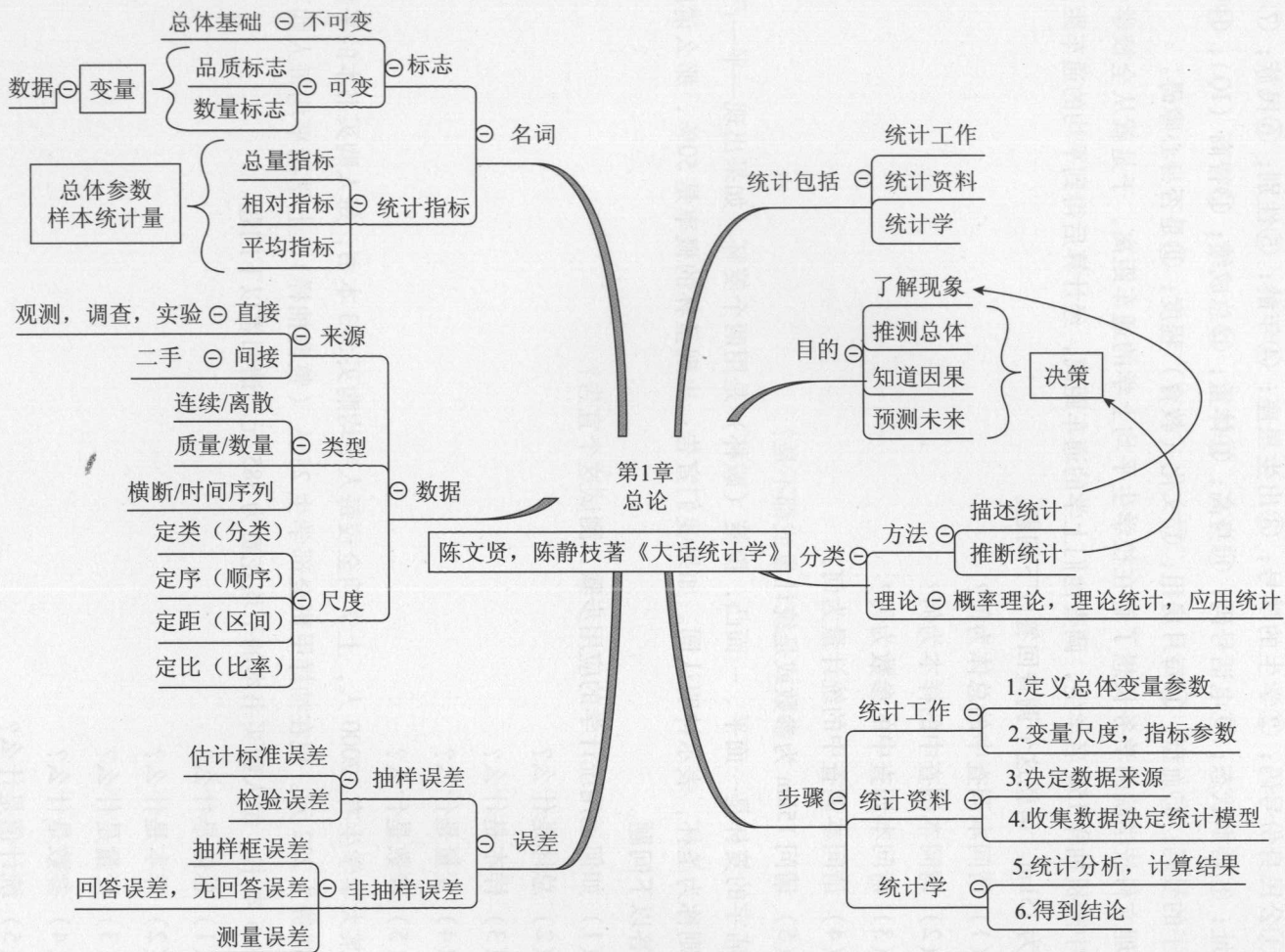


图1-8 第1章思维导图

习 题

1. 说明下列数据变量的衡量尺度和数据变量的分类：
①公民身份号码；②学生的学号；③出生星座；④年龄；⑤性别；⑥民族；⑦住址；⑧婚姻状态；⑨电话号码；⑩身高；⑪体重；⑫总成绩；⑬智商（IQ）；⑭鞋子的号码；⑮血型；⑯每月房租；⑰文化（教育）程度；⑱是否身心障碍。
2. 理工科技学院的学务长想了解在校学生平均上学的通车距离，于是就从全校学生中随机抽样 150 名学生，调查他们上学的通车距离，经计算后得到平均的通车距离为 15km。请就这个调查回答以下问题。
 - (1) 请问本调查中的总体为何？
 - (2) 请问本调查中的样本为何？
 - (3) 请问本调查中的参数为何？
 - (4) 请问本调查中的统计量为何？
 - (5) 请问 15km 为参数或是统计量或都不是？
3. 庙宇的筊杯是一面平、一面凸，掷筊（跋杯）是用两个筊杯，如果出现一平一凸，则称为圣杯，表示神明认同。如果我们宣告：出现圣杯的概率是 50%，那么请回答以下问题。
 - (1) 如何运用统计学的应用步骤，测试这个宣告？
 - (2) 总体是什么？
 - (3) 样本是什么？
 - (4) 变量是什么？
 - (5) 参数是什么？
4. 某大学学生有 12000 人，上学期全校每人平均购买 5.3 本书，每人购买书本的平均花费为 232 元。现在抽样甲班全部学生 25 人（整群抽样），上学期平均每人购买 8.5 本书，每人购买书本的平均花费为 385 元。请回答以下问题。
 - (1) 总体是什么？
 - (2) 样本是什么？
 - (3) 变量是什么？
 - (4) 参数是什么？
 - (5) 统计值是什么？

(6) 对于甲班学生你有什么看法?

(7) 如果甲班学生平均每人购买 8.5 本书, 平均花费为 1265 元, 那么你的看法是什么?

其他习题请下载。



第2章

描述统计

巧匠为宫室，为圆必以规，为方必以矩，为平直必以准绳。

——《吕氏春秋·分职》

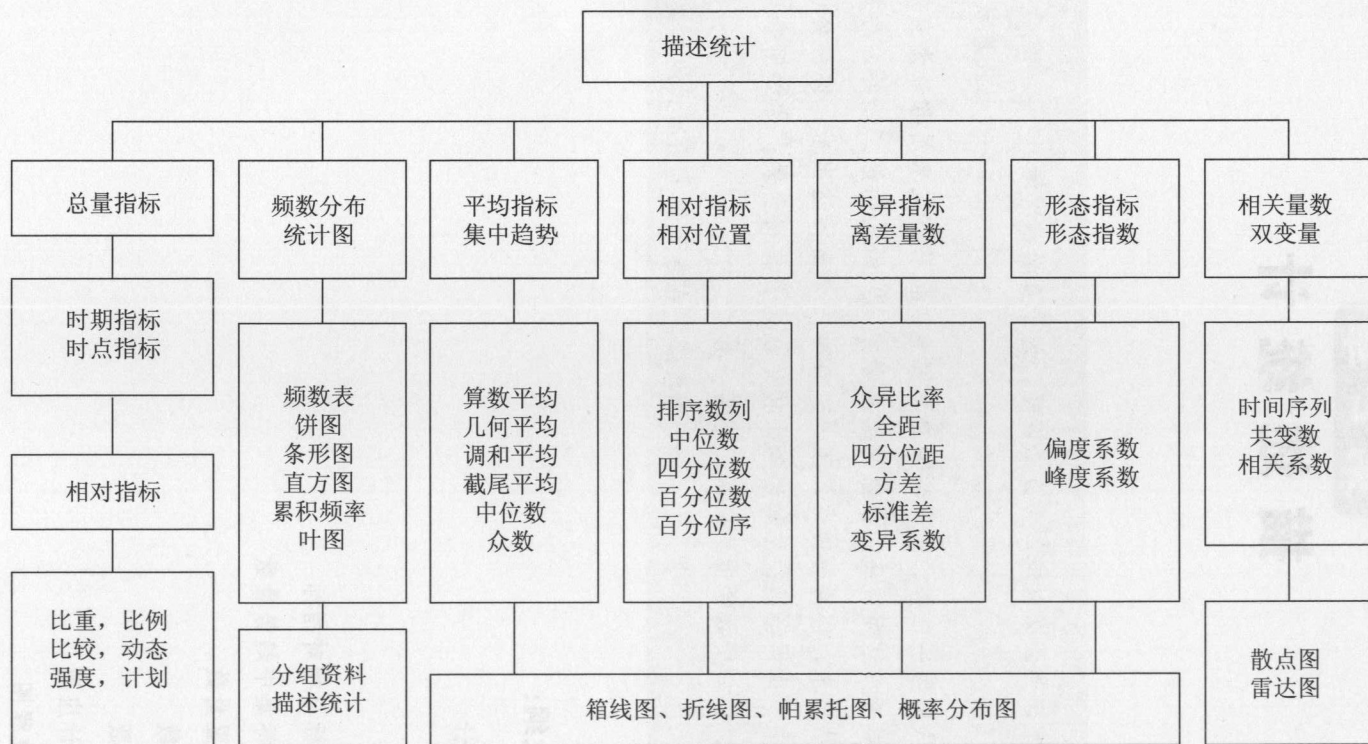
所以，做人的第一任务，是发现你自己究竟以何种方式来思考最为有效；你究竟是图画式的思考方法呢，还是抽象式的思考方法？做人的第二任务，乃是为了把你所思考的传达出来，你得有传达的工具。你如用语言传达出来，你得会说；你如用图画传达出来，你得会画；你如用舞蹈传达出来，你得会跳；你如用音乐传达出来，你得会唱；你如用数学传达出来，你得会算。

——陈之藩《在春风里·图画式的与逻辑式的》



本章重点大纲：

- 2.1 描述统计
- 2.2 统计表
- 2.3 统计图
- 2.4 总量指标与相对指标
- 2.5 平均指标集中趋势量数
- 2.6 相对位置量数
- 2.7 离差量数
- 2.8 形态量数
- 2.9 中文统计应用
- 2.10 本章流程图
- 2.11 本章思维导图



本章概念图

2.1 描述统计

描述统计是将收集或调查得到的数据，加以整理或简化，将其表达成较有意义的信息数值或图形。描述统计描述数据的方法有：图形法、集中趋势量数、离差量数、位置量数、形态量数。离差量数又被翻译为离散量数，但是离差量数有“离散”和“差异”的意思，还有这里的“离散”和随机变量型态的“离散型”（discerte）意义不同。

集中趋势量数又称作代表值量数（representative measures），最主要的集中趋势量数是（算数）平均数；离差量数是衡量数据的分散变化，最主要的离差量数是方差或标准差。平均数和方差，可以简单地描述总体的分布。

同样的平均数，有人喜欢方差大（风险刺激），有人喜欢方差小（稳定保守）。例如：股票投资的风险追求者（富贵险中求）或风险避免者（平安是福）；居住地区，有人喜爱四季如春（温度方差小）、有人偏好四季分明（温度方差大），都有不同的选择。

2.2 统计表

2.2.1 分类统计表的构造

分类统计表的结构包括：总标题、横行标题、纵栏标题和统计数据，必要时可以在表的下方加上表外附加或附注，如表 2-1 所示。

表2-1 国家统计局 2013年GDP（国内生产总值）

产业		绝对值(亿元)	比上年同期增长(%)
横 行 标 题	第一产业	56957	4.0
	第二产业	249684	7.8
	第三产业	262204	8.3
	GDP	568845	7.7

主词

宾词

← 总标题

← 纵栏标题

← 统计数据

2.2.2 频数分布表

频数分布（frequency distribution），通常是将数据整理成一个表，有下列 3 种数据类型：①单值分组频数分布表；②离散型数据频数分布表；③连续型数据频数分布表。

（1）单值分组（single - value grouping）频数分布表，又称作列举式（list）分组，适合于定类或定序尺度数据，其统计图不适宜用直方图，可用条形图，如表 2-2 所示。

表 2-2 分组数据频数分布表

组界	频数	频率
0	12	0.400
1	6	0.200
2	7	0.233
3	3	0.100
4	2	0.067
	30	1.000

(2) 离散型数据频数分布表：离散型数据是整数，没有小数或分数，如表 2-3 所示。

表 2-3 离散型数据频数分布表

组界	组中点	频数	相对频数	(向上) 累积频数	直方图组界
20 ~ 29	24.5	10	0.40	10	19.5 ~ 29.5
30 ~ 39	34.5	7	0.28	17	29.5 ~ 39.5
40 ~ 49	44.5	4	0.16	21	39.5 ~ 49.5
50 ~ 59	54.5	2	0.08	23	49.5 ~ 59.5
60 ~ 69	64.5	2	0.08	25 上	59.5 ~ 69.5
		25	1.00		

(3) 连续型数据频数分布表有两种型式：第一种型式，下组界（lower class limit）属于该组，但是上组界（upper class limit）不属于该组；第二种型式，下组界不属于该组，但是上组界属于该组。频数分布表多采用第一种型式，如表 2-4 所示。中文统计的频数分布表也是第一种型式。

表 2-4 连续型数据频数分布表

组界	组中点	频数
[20, 30)	25	10
[30, 40)	35	7
[40, 50)	45	4
[50, 60)	55	2
[60, 70)	65	2
		25

频数分布表的制作步骤如下。

- (1) 将数据按照大小顺序排列数列。
- (2) 选择数据的衡量尺度或形态, 选择频数分布表的型式(离散型、连续型或单值型)。
- (3) 定出数据的范围(range) $R = \text{最大值} - \text{最小值}$, 或者为了配合上下限再增大范围。

(4) 选择分组的数目 k 。如果数据的数目是 n , 则建议分组的数目为: $k = 3.3(\log_{10} n) + 1$ 。

例如: $n = 100, k = 3.3(\log_{10} n) + 1 = 3.3(2) + 1 = 7.6, k \approx 8$;

若 $n = 25$, 则 $k = 3.3(\log_{10} n) + 1 = 3.3(1.4) + 1 = 5.6, k \approx 6$, 建议分为 6 组。

(5) 计算每组区间的长度, 称作组宽(class width)或组距(class interval)。通常组宽是最接近 R/k , 而且较容易计算的数值, 如 5, 10, 100。

(6) 选出每组的下组界和上组界: 使用简单的组界(例如: 避免组界为 17.394)。

(7) 计算每组的出现频数(frequencies): 每个数据一定属于一个组, 而且只属于一个组, 称为归类。再利用划记, 如“正”或“-”的记号, 计算频数。

(8) 计算每组的频率(relative frequencies, 频率百分比): 出现频数除以数据总数。

(9) 必要时, 可在频数分布表再加两栏, 一栏是第一组到各组的累积频数(cumulative frequency), 另一栏是累积频率(cumulative relative frequency)。累积频数通常用“向上累积”, 往数据大的组累积。

频数分布表注意事项如下。

(1) 每组是“互斥”的, 没有共同的数据; 所有组是“遍及”(完备)全部数据的。换句话说, 每个数据一定会属于一组, 而且只属于一组。

(2) 每组组宽尽量相等, 但是如果数据的分布不好, 可能会合并某些组, 使得组宽不等。在画直方图时, 组宽不同要修改频数, 请见 2.3.1 节直方图。

(3) 最好避免开放组界(open-ended class), 即上组界或下组界是无限。但是如果连续有很多组的频数为“0”, 且数据分布很散, 则可能有开放组界。

例题 2.1 学生的成绩频数分布表(数据请见例题 2.2, 解答请见网络资源)

2.3 统计图

统计图包括: 直方图、多边形图、累积频率图、饼图、折线图、条形图、茎叶图、箱线图、帕累托图等。大部分图形都是与叙述数据分布有关。折线图反映时间的变化,

直方图、茎叶图都是叙述分组频数分布的图形。选择统计图时要注意适合什么尺度的数据。

2.3.1 直方图

直方图 (histogram) 是利用频数分布，将每组频数，以 X 轴为组界， Y 轴为频数或频率，画出长方形的图形，如图 2-1 所示。

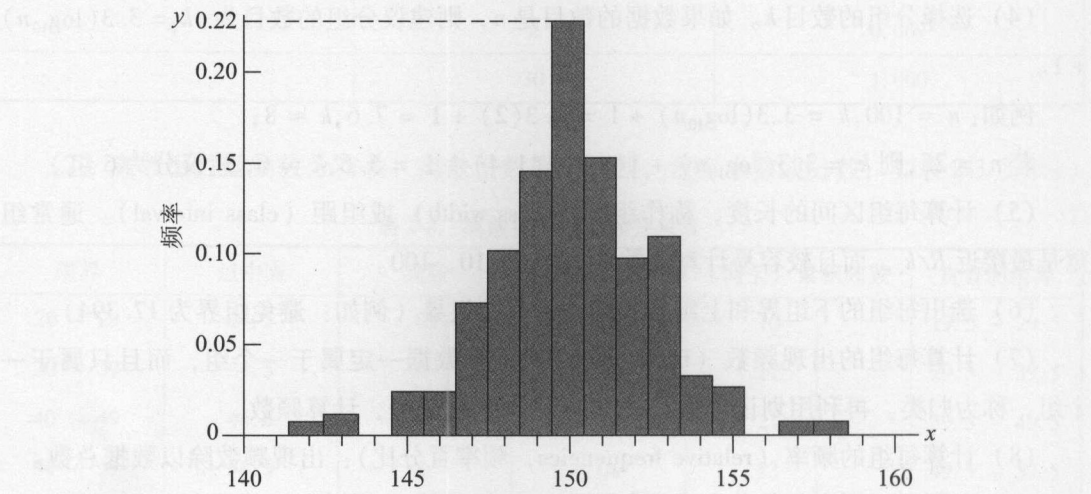


图 2-1 直方图

由直方图可以大致知道数据是否符合或近似正态分布（对称与正态形状），图 2-2 就不符合正态分布。

如果频数分布表的组宽不相等，要画出直方图，则比较长的组宽的频数和频率要加以修正降低。组宽 2 倍的频数和频率要除以 2；组宽 5 倍的频数和频率要除以 5。

背对背直方图 (back-to-back histogram) 可以比较两组数据的直方图，即一个变量（人口数）两组总体（男女）数据的比较，如图 2-2 所示。

例题 2.2 学生的成绩直方图。（解答见网络资源）

25, 32, 35, 35, 35, 36, 40, 42, 44, 46, 47, 48, 48, 49, 50
50, 57, 58, 66, 72, 78, 85, 86, 87, 88, 89, 91, 92, 94, 95

例题 2.3 不同组距频数分布表，画直方图。（解答见网络资源）

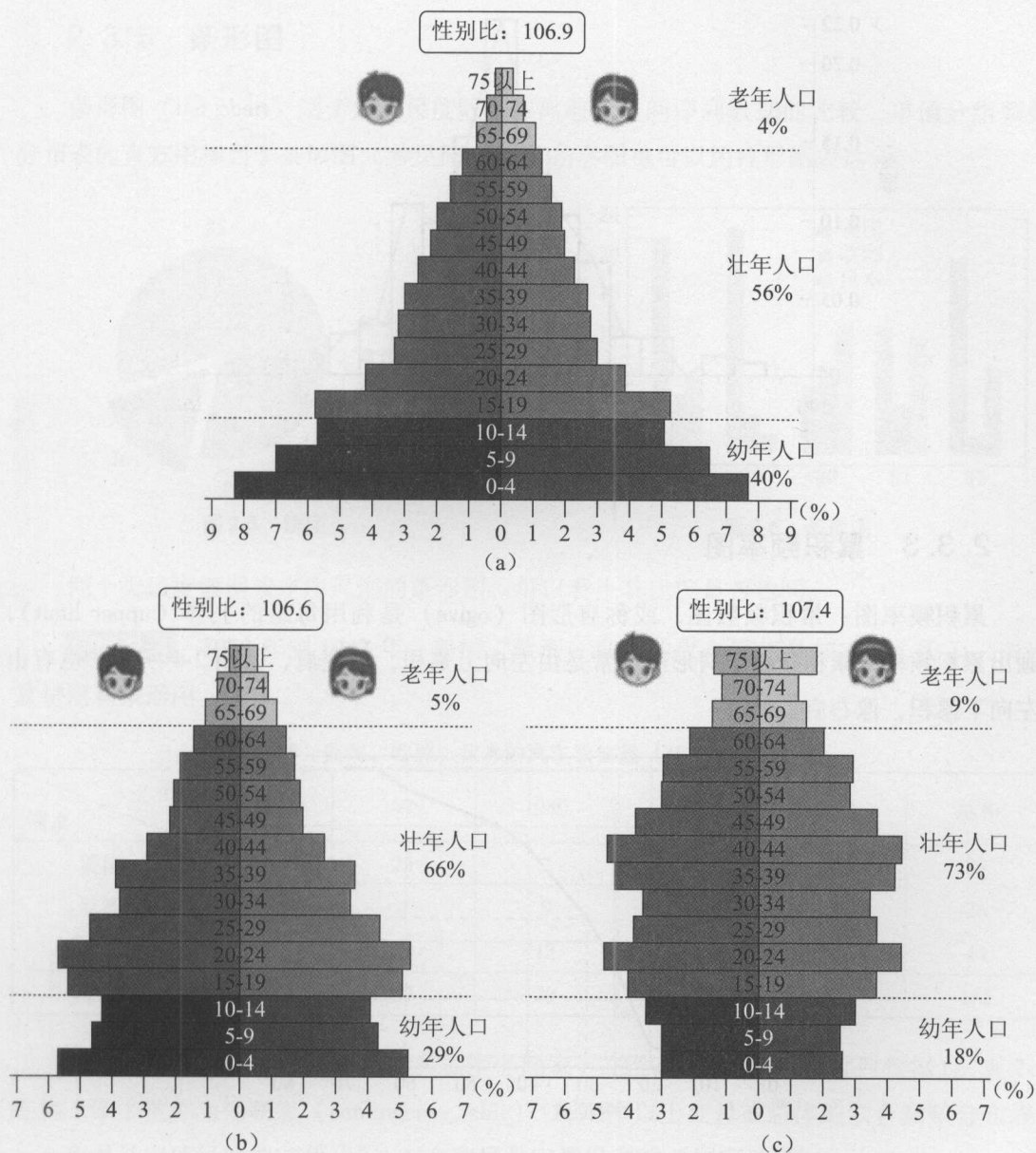


图 2-2 中国人口金字塔 (年龄和性别的结构) (中文参考书目 [20])

(a) 1970 年; (b) 1990 年; (c) 2010 年

2.3.2 多边形图

多边形图 (polygon) 是将直方图的组中点 (midpoint), 即每组中点的频数, 连接起来, 描出多边形的图形如图 2-3 所示。可用于比较两组数据。

多边形图比直方图更容易看出数据分布的形态, 例如是否正态分布, 偏度或峰度等。

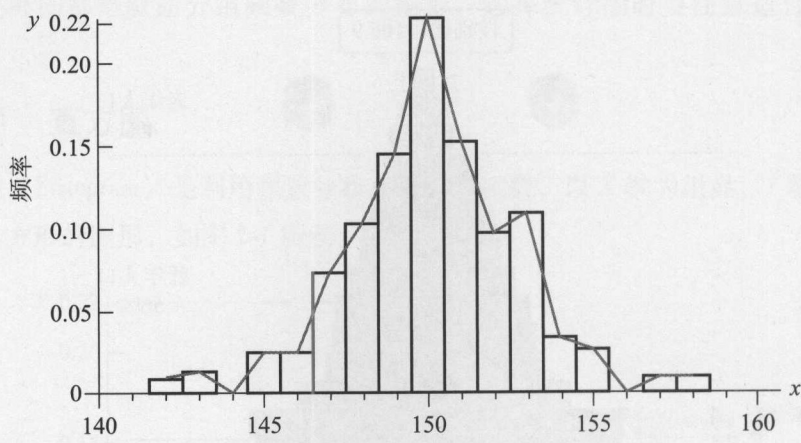


图 2-3 多边形图

2.3.3 累积频率图

累积频率图、累积频数图，或称肩形图（ogive）是利用每组的上界（upper limit），画出累积频率或累积频数。肩形图通常是由左向上累积，像左肩，如图 2-4 所示；也有由左向下累积，像右肩。

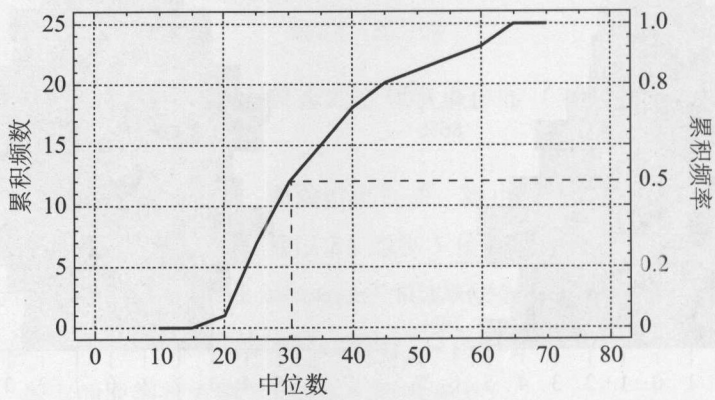


图 2-4 累积频数（频率）图

累积频率图的纵坐标是频率 0~1，在 0.5 往右画虚线，遇到累积频率图的图形，往下画虚线，遇到横坐标，即为中位数。图 2-4 的中位数是 30。

2.3.4 饼图

饼图（pie chart）适于定类尺度数据或定序尺度数据的比较（参见图 2-5）。

2.3.5 条形图

条形图 (bar chart) 适于定类尺度数据或离散型时间序列数据的比较。单值分组频数分布表的直方图相当于条形图 (参见图 2-6)。条形图也可以用柱形图表达。

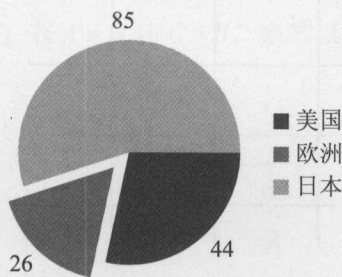


图 2-5 饼图

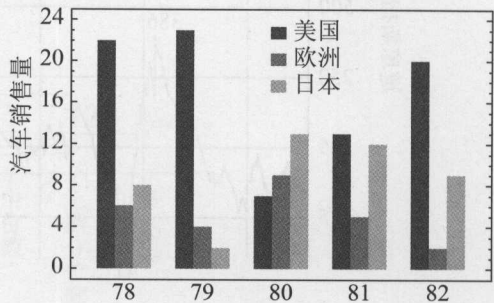


图 2-6 条形图

两个类尺度数据或定序尺度的条形图，可以看出其比例是否相同。

例题 2.4 1978 年到 1982 年，美国、欧洲、日本的汽车销售量如表 2-5 所示，请计算饼图和条形图。

表 2-5 美国、欧洲、日本的汽车销售量 (1978—1982 年)

国家 \ 年	1978	1979	1980	1981	1982	总和
美国	22	23	7	13	20	85
欧洲	6	4	9	5	2	26
日本	8	2	13	12	9	44
总和	36	29	29	30	31	155

请注意这是单变量 (销售量) 多总体 (国家) 的纵向数据 (时间数列数据)。表 2-5 在第 5 章中被称为列联表 (contingency table)，即两个以上变量或总体的联合频数分布表，一个变量是定序尺度的“年代”，一个变量是定类尺度的“国家地区”。

解答：

图 2-6 的中文统计应用请见 2.9.4 小节。

2.3.6 折线图

折线图 (line chart) 适用于时间序列 (time series) 数据的比较，如图 2-7 所示。

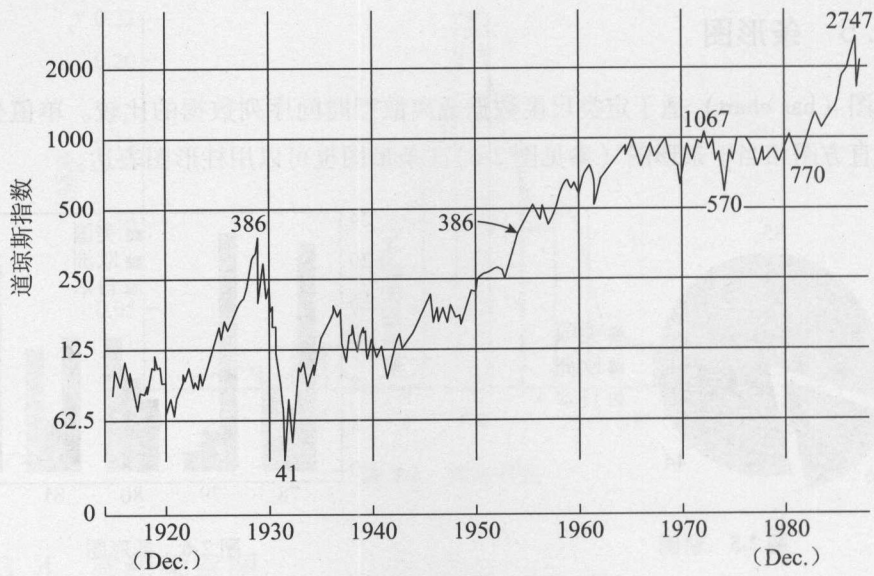


图 2-7 美国道琼斯指数的折线

注：Y 轴的尺度不相等，上面的变动幅度应较大。

2.3.7 茎叶图

在直方图中，长方形代表实际数字。茎叶图（stem and leaf plot）混合了数字与图形，适用于定比尺度和定距尺度的离散数据，如图 2-8 所示。茎叶图逆时针旋转 90°，可以看作直方图。

4	2	->5
5	3	->25556
6	4	->02467889
7	5	->0078
8	6	->6
9	7	->28
10	8	->56789
11	9	->1245

图 2-8 学生的成绩茎叶图

2.3.8 箱线图

箱线图（box-and-whisker plot）或称方盒图（box plots）画出数据的全距、四分位数、中位数等数值，其计算请见下文。多组箱线图（multiple box-and-whisker plot）可以比较几

组数据。由箱线图可以看出数据分布的形态：对称型分布、右偏分布、左偏分布，如图 2-9 所示。

- (1) 若 $A = D$ 且 $B = C$ ，则为对称。
- (2) 若 $A > D$ 且 $B > C$ ，则为左偏。
- (3) 若 $D > A$ 且 $C > B$ ，则为右偏。
- (4) 若 $A > D$ 且 $C > B$ ，则难以判定偏态。

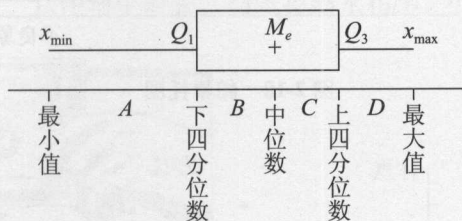


图 2-9 箱线图

有关左偏、右偏，请见 2.8.2 节。

例题 2.5 学生的成绩箱线图。(解答见网络资源)

2.3.9 帕累托图

帕累托或称柏拉图 (Vifredo Pareto, 1848—1923) 是意大利经济学家，在分析财富的分布时，发现少数人 (20%) 拥有多数的财富 (80%)，这就是 80/20 原则。帕累托图 (Perato chart) 是按照频数由大到小排列的直方图，其累积频数曲线就是帕累托曲线，如图 2-10 所示。帕累托图通常应用在质量管理，它结合了直方图与累积频率图，表示某些原因造成不良品的频数与累积概率，频数分布是由大到小排列的。以累计至 80% 的前数项原因为改善对象，因为 20% 的原因可能造成 80% 的不良数目。若帕累托曲线近似直线，则贫富差距缩小；帕累托曲线越弯曲，表示的贫富差距越大，可能是 90/10 原则，10% 的产品占有 90% 的营业额。帕累托图可应用在物料管理重点管理的 ABC 分析，A 类物料就是数量 20%，价值 80%。

例题 2.6 帕累托图。(解答见网络资源)

2.3.10 散点图

散点图即两个变量的分布图，如图 2-11 所示。

例题 2.7 散点图。(解答见网络资源)

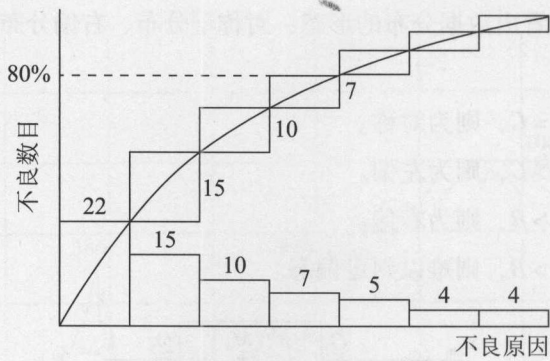


图 2-10 帕累托图

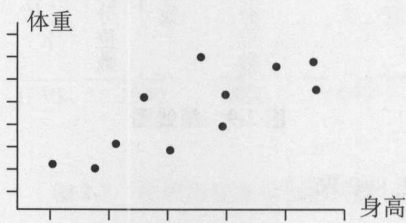


图 2-11 散点图

2.3.11 雷达图

雷达图是表示多个变量或指标的分布图，如图 2-12 所示。

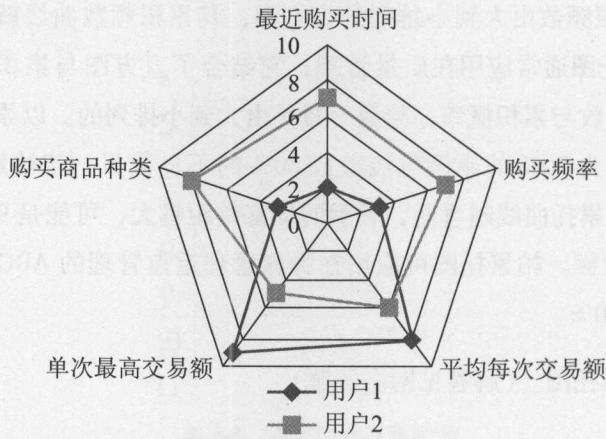


图 2-12 雷达图（中文参考书目 [19]）

例题 2.8 雷达图。（解答见网络资源）

2.3.12 统计图的应用

利用统计图，要注意“欺骗因素”（lie factor）。所谓欺骗因素是指如果用二度空间或三度空间来表达统计图，且只用一边（一个纬度）来表示比率，那么结果因为二度空间的关系，面积的比率是原来的比率的平方。

例如图 2-13，1978 年，1 美元的购买力相当于 1958 年的 0.44 美元，但是图形的面积比率是 $(0.44)^2 = 0.1036$ ，从图形上看来，会觉得缩水很多。

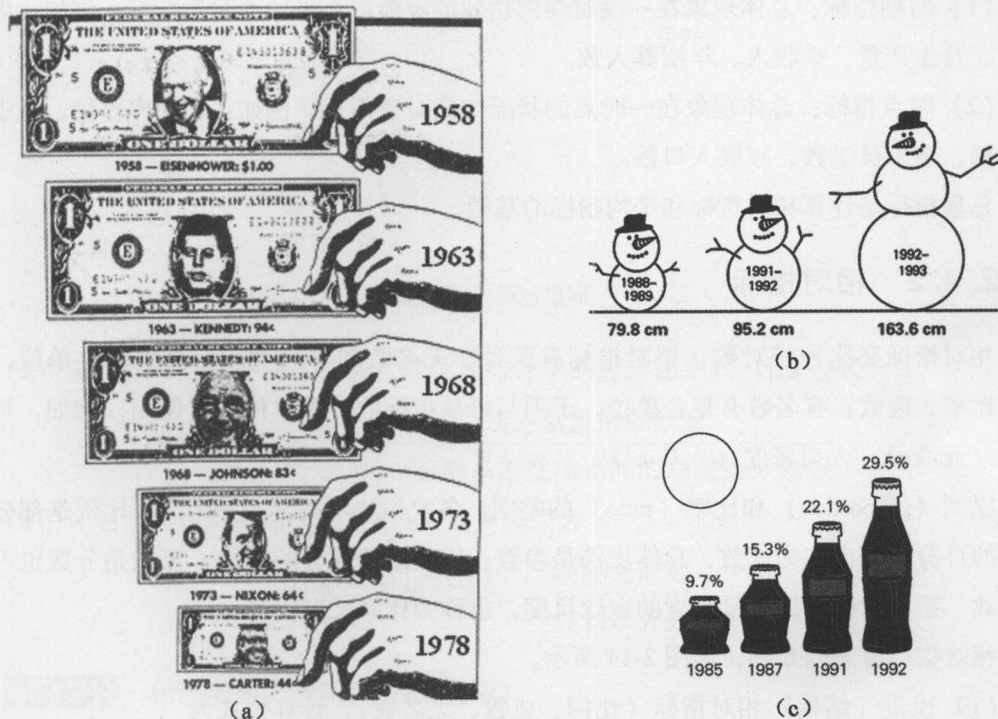


图 2-13 统计图的欺骗因素

(a) 1 美元的购买力；(b) 下雪厚度；(c) 可乐报酬率

下雪厚度的统计图也是有欺骗因素。可乐报酬率的统计图是可以 OK 的，因为其底部是相同的。

Excel 的图形，如圆柱图、圆锥图、金字塔图，虽然是 2D 或 3D 立体图，但是底部相同，所以没有欺骗因素。

2.4 总量指标与相对指标

2.4.1 总量指标

总量指标是反映现象在一定时间、地点和条件下的总数量、总水平、总规模的统计指标，又称为绝对数指标。绝对数指标有以下两类。

(1) 时期指标：总体现象在一段时期内达到的规模或水平的绝对数指标。例如：周销售额、月生产量、季收入、年招募人数。

(2) 时点指标：总体现象在一时点的状况的绝对数指标。例如：周初库存量、月中存款余额、季末员工数、年底人口数。

总量指标是计算相对指标和平均指标的基础。

2.4.2 相对指标

相对指标又称为相对数，相对指标有两种：无名数和有名数。无名数没有单位，例如，比率、成数；有名数有复合单位，子项与母项指标的计量单位同时使用，例如，单位成本（元/kg）、人口密度（人/km²）。

比例（proportion）和比率（ratio）的差别，各家说法不一。本书定义：比例是部分占全部的百分比，也称为成数，总体比例是参数；比率是两部分的对比，可能是分数也可能是倍数，第1章中数据衡量尺度的定比尺度，也称为比率尺度。

相对指标有下列6种，如图2-14所示。

(1) 比重（结构）相对指标（比例、成数、无名数），计算公式为

$$\text{比重(结构)相对指标} = \frac{\text{总体中某部分值}}{\text{总体数值}}$$

(2) 比率相对指标（无名数），计算公式为

$$\text{比率相对指标} = \frac{\text{总体中某一部分值}}{\text{总体中另一部分值}}$$

(3) 比较相对指标（无名数），计算公式为

$$\text{比较相对指标} = \frac{\text{某空间的某指标值}}{\text{某空间的某指标值}}$$

(4) 动态相对指标（指数、发展速度、无名数），计算公式为

$$\text{动态相对指标} = \frac{\text{报告期水平}}{\text{基期水平}}$$



图 2-14 相对指标

(5) 强度相对指标（有名数或无名数如会计学的速动比率），计算公式为

$$\text{强度相对指标} = \frac{\text{某一总量指标数值}}{\text{另一有联系但性质不同的总量指标数值}}$$

(6) 计划完成相对指标（无名数），计算公式为

$$\text{计划完成相对指标} = \frac{\text{实际完成数}}{\text{计划任务数}}$$

例题 2.9 相对指标。（解答见网络资源）

2.5 平均指标集中趋势量数

集中趋势量数是计算出代表全部数据的数值，例如：平均数、中位数、众数等。下列各节将分别进行说明。

集中趋势量数和相对位置量数的主要功能如下。

- (1) 简化一组数据，将全部数据简化为数个数据。
- (2) 代表数据的集中程度，表示数据的集中或重心。
- (3) 比较两组以上的数据，表示两组总体数据的差别。

2.5.1 算术平均数

定义 总体数据 x_1, x_2, \dots, x_N 等 N 个数据（区间或定比尺度），或样本数据 x_1, x_2, \dots, x_n 等 n 个数据，定义算术平均数（arithmetic mean），简称平均数或均值（mean）。

（1）总体平均数，计算公式为

$$\mu = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N}$$

（2）样本平均数，计算公式为

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

平均数是最重要的代表值，是集中趋势的最主要指标，因为它是数据的平衡点。所谓数据的平衡点，就是平均数左边（小于平均数）的数据到平均数的距离总和，等于平均数右边（大于平均数）的数据到平均数的距离总和，如图 2-15 所示。

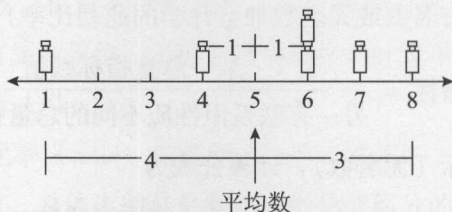


图 2-15 数据的平衡点

算术平均数的特点如下。

（1）常识性代表值，数值重心。每个数据与平均数之差的总和为 0。

$$\sum_{i=1}^N (x_i - \mu) = \sum_{i=1}^n (x_i - \bar{x}) = 0$$

（2）推断统计的主角（检定与估计的主要参数）。

（3）比较两组以上的数据。

（4）能进行代数处理，已知数组的算术平均数，可直接计算其总平均数。例如，有 m 组样本数据，每组有 n_i 个数据，平均数为 \bar{x}_i ，则总平均数 \bar{x} 为

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \dots + n_m x_m}{n_1 + n_2 + \dots + n_m} = \sum_{i=1}^m n_i x_i / \sum_{i=1}^m n_i$$

（5）代表所有数据，敏感度高（若有一个数据改变，则平均数也会改变），因为每个数据值都用来计算。

(6) 各数值与平均数之差的平方和最小, 即每个数据到平均数的距离的平方加起来是最小的, 即

$$\sum_{i=1}^N (x_i - \mu)^2 \leq \sum_{i=1}^N (x_i - r)^2, \forall r \in R \quad \text{或} \quad \sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - r)^2, \forall r \in R$$

证明:

$$\begin{aligned} \sum (x_i - r)^2 &= \sum [(x_i - \mu) + (\mu - r)]^2 \\ &= \sum (x_i - \mu)^2 + 2 \sum (x_i - \mu)(\mu - r) + \sum (\mu - r)^2 \\ &= \sum (x_i - \mu)^2 + 2(\mu - r) \sum (x_i - \mu) + \sum (\mu - r)^2 \\ &= \sum (x_i - \mu)^2 + n(\mu - r)^2 \geq \sum (x_i - \mu)^2 \end{aligned}$$

(7) 若一组数据 (不管是总体或抽样) x_i , 经过线性转换

$$y_i = a + bx_i, \forall a, b \in R$$

则 x_i 的算数平均数 μ_x 与 y_i 的算术平均数 μ_y 的关系

$$\mu_y = a + b\mu_x$$

(8) 各平均数的大小关系: 算术平均数大于几何平均数, 几何平均数大于调合平均数, 即

$$\mu \geq \mu_G \geq \mu_H \quad \text{或} \quad \bar{x} \geq \bar{x}_G \geq \bar{x}_H$$

(9) 算术平均数、几何平均数、调合平均数, 各有其适用的数据类型。

关于几何平均数、调合平均数后面有讲解。

例题 2.10 学生的成绩平均数。(解答见网络资源)

2.5.2 加权平均数

算术平均数, 假设所有的数据有相同的重要性或权重, 实际上有的平均数要考虑每个数据的权重, 例如大学生每学期的总平均成绩, 以所修各科目的学分数为权重; 股票市场的加权指数, 以资本额为权重; 企业选择人才、计算机设备、工厂地点、投资股票、绩效指标, 要先决定衡量因素及其权重, 再根据分数计算加权平均。

定义 假设一组样本数据 x_1, x_2, \dots, x_n 等 n 个数据, 其对应权重为 w_1, w_2, \dots, w_n , 则加权平均数定义如下

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{w_1 + w_2 + \dots + w_n} = \sum_{i=1}^n x_i / \sum_{i=1}^n w_i$$

2.5.3 几何平均数

定义 假设一群数字数据有 x_1, x_2, \dots, x_n 等 n 个数据, 几何平均数 (geometric mean) 记作 G 或 \bar{x}_G, μ_G , 其定义如下

$$G = \mu_G = \bar{x}_G = \sqrt[n]{x_1 \times x_2 \times \dots \times x_n} = \sqrt[n]{\prod_{i=1}^n x_i}$$

计算指数, 几何平均是理想公式。几何平均数适用于“动态相对数”(发展速度)的数据(例如: 增长率、指数等) r_i , 其平均成长率为

$$G - 1 = \sqrt[n]{(1 + r_1) \times (1 + r_2) \times \dots \times (1 + r_n)} - 1$$

例如: 连续两年, 第一年成长为 2 倍 (成长率 100%), 第二年成长为 8 倍 (成长率 700%), 平均每年成长率为 $\sqrt[2]{(1+1) \times (1+7)} - 1 = 300\%$, 而不是平均每年成长 $(1+7)/2 = 400\%$ 。

几何平均数通常应用在无名数动态相对指标, 例如第 3 章的发展速度。

例题 2.11 学生的成绩几何平均数。(解答见网络资源)

2.5.4 调和平均数

定义 数据 x_1, x_2, \dots, x_n 等 n 个数据, 调和平均数 (harmonic mean) 记作 H 或 \bar{x}_H, μ_H , 其定义如下

$$H = \mu_H = \bar{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

调和平均数适用于“强度相对数”的数据, 例如: 速率 (每小时多少千米)、物价 (每件多少钱), 而且“各数据的总值相等”(总千米, 总价相等)的数值数据求平均。

如果“各数据的总值 (标志总量) M_i 不相等”, 要计算“各单位数据 x_i 平均”, 而“全部数据总值” = $\sum M_i$, M_i 是加权调和平均数的权数, 则加权调和平均数的定义如下

$$H = \frac{\text{全部数据总值}}{\sum \frac{\text{各数据的总值}}{\text{各单位数据}}} = \frac{\sum M_i}{\sum \frac{M_i}{x_i}}$$

若 M_i 都相等, 则这个式子等于调和平均数。

调和平均数通常应用在有名数强度相对指标。

加权调和平均数 (已知 x_i, M_i) 等于加权算术平均数 (已知 x_i, f_i), 即

$$H = \frac{\sum \frac{M_i}{x_i}}{\sum \frac{M_i}{x_i}} = \frac{\sum x_i f_i}{\sum f_i} = \bar{x}$$

表 2-6 加权调和平均数（权数 M_i ）与加权算术平均数（权数 f_i ）

总体单位 i	变量值 x_i 强度指标（单位成本）	单位数 f_i	标志总量 M_i （总额）
1	x_1	f_1	M_1
2	x_2	f_2	M_2
\vdots	\vdots	\vdots	\vdots
n	x_n	f_n	M_n
总和	$\sum x_i$	$\sum f_i$	$\sum M_i$
$x_i \times f_i = M_i$			

例题 2.12 学生的成绩调和平均数。（解答见网络资源）

例题 2.13 定期定额投资。（解答见网络资源）

例题 2.14 怎么少了 100 元？

两名男子在卖领带：老王是两条卖 100 元，老李是 3 条卖 100 元。他们决定以合作代替竞争，两人各拿出 30 条，凑成 60 条，并且决定每条卖 40 元，因为“两条 100 元”加上“3 条 100 元”等于“5 条 200 元”即“每条 40 元”。于是两人合作以每条 40 元卖掉 60 条，得到 2400 元。但是，如果老王以两条 100 元卖掉 30 条，可以得到 1500 元；老李以 3 条 100 元卖掉 30 条，可以得到 1000 元，总和应该是 2500 元。请问：怎么少了 100 元？如果老王拿出 30 条，老李拿出 90 条，那么应如何定价？

解答：“两条 100 元”是每条 50 元，“3 条 100 元”是每条 33.33 元。
其算术平均数为 $(50 + 33.33) / 2 = 41.66$ ，每条 41.66 元。
其调合平均数为 $2 / (1/50 + 1/33.33) = 40$ ，每条 40 元（5 条 200 元）。
如果以“5 条 200 元”来卖，则老王应该拿出 24 条（价值 1200 元）；老李应该拿出 36 条（价值 1200 元），总值相同。总共 60 条卖得 2400 元；然后各分 1200 元。
如果两人各拿出 30 条，凑成 60 条，则每条应该卖 41.66 元（3 条 125 元），总共得到 2500 元；然后老王分 1500 元，老李分 1000 元。

所以，利用算术平均数应注意其“数量相等”（数量不相等，以数量加权），利用调合平均数应注意其“总值相等”（总值不相等，以总值加权）。
如果老王拿出 30 条，老李拿出 90 条，应用加权平均数定价，即

$$(50 \times 30 + 33.33 \times 90) / 120 = 37.5 \quad (\text{两条 } 75 \text{ 元})$$

请注意：“算术平均数大于调合平均数，几何平均数 (40.82) 则介于其中间”。

2.5.5 中位数

中位数可以是集中趋势量数，也可以是相对位置量数。

定义 假设一群数字数据有 x_1, x_2, \dots, x_n 等 n 个数据，已按照大小，由小到大排列，即 $x_1 \leq x_2 \leq \dots \leq x_n$ 。中位数 (median)，记作 M_e ，是有一半的观察值在 M_e 之前，另有一半的观察值在 M_e 之后。

定义 如果以数据的个数来计算中位数，则

$$M_e = \begin{cases} x_{(n+1)/2}, & \text{若 } n \text{ 是奇} \\ \frac{x_{n/2} + x_{(n/2)+1}}{2}, & \text{若 } n \text{ 是偶} \end{cases}$$

(1) 当数据中有极端值的改变，即有非常大或非常小的数值，则平均数会有很敏感的改变，但是中位数则不受影响。例如：一个公司员工的平均薪资所得，如果 5 个样本中有一个年所得数千万的总经理；一个城市的房屋平均成交价格，如果 6 个样本中有一个价值数亿的豪宅；一个学生班级的平均成绩，如果 4 个样本中有一个 0 分的学生。

(2) 各数值与中位数之差的绝对值之和为最小，即

$$\sum_{i=1}^n |x_i - M_e| \leq \sum_{i=1}^n |x_i - r| \quad \forall r \in \mathbb{R}$$

(3) 当数据是定序尺度时，集中趋势不适合以平均数为代表值，应该以中位数为代表值。

例题 2.15 学生的成绩中位数。(解答见网络资源)

2.5.6 众数

定义 假设一群数字数据有 x_1, x_2, \dots, x_n 等 n 个数据，众数 (mode) 记作 M_o ，定义如下

$$M_o = \text{出现频数最多的 } x_i \text{ 值}$$

有的数据，有两个众数，称作双众数 (bimodal)；有两个以上众数，称作多众数 (multimodal)；如果每个数据都只出现一次或都有相同频数，则称作无众数。

例题 2.16 学生的成绩众数。(解答见网络资源)

2.5.7 截尾平均数

有时候,数据中最大的几个数和最小的几个数,是异常现象,会影响到平均数的代表性。所以,将首尾两端的数据去掉,再取其平均数,称之为截尾平均数(trimmed mean)。利用裁判评分的比赛,如体操、选美等,由于评分可能不客观,所以选择截尾平均数,决定其名次,以避免有一两位评审不公平。

定义 假设一群数字数据有 x_1, x_2, \dots, x_n 等 n 个数据,已按照大小,由小到大排列,即 $x_1 \leq x_2 \leq \dots \leq x_n$, $p\%$ 截尾平均数是首尾各去掉 $p\%$ 再作平均数,记作 \bar{x}_T , 即

$$\bar{x}_T = \frac{\sum_{i=k+1}^{n-k} x_i}{n-2k}$$

式中: $k = \left\lfloor \frac{pn}{100} \right\rfloor$ 取整的部分, 即 $k = \left\lfloor \frac{pn}{100} \right\rfloor$ 。

定义 假设一群数字数据有 x_1, x_2, \dots, x_n 等 n 个数据,已按照大小,由小到大排列。温瑟平均数 W (Winsorized mean) 是将下四分位数 Q_1 以下的数,改为 Q_1 ; 将上四分位数 Q_3 以上的数,改为 Q_3 ; 再将这 n 个数据计算算术平均数,即

$$W = \frac{k(Q_1 + Q_3) + \sum_{i=k+1}^{n-k} x_i}{n}$$

式中: $k = \left\lfloor \frac{n}{4} \right\rfloor$ 取整的部分, 即 $k = \left\lfloor \frac{n}{4} \right\rfloor$ 。

例题 2.17 学生的成绩温瑟平均数。(解答见网络资源)

例题 2.18 奥运双人跳水分数的计算。运动员的分数如图 2-16 所示。



GBR  BLAKE ALDRIDGE / THOMAS DALEY									
ROUND 1					TOTAL 52.80				
Difficulty 2.0					Score 52.80				
EX1	EX2	EX3	EX4	SY1	SY2	SY3	SY4	SY5	
8.5	8.5	8.5	8.5	9.0	9.0	9.0	9.0	9.0	

图 2-16 运动员的分数

- 解答: (1) 执行 (EX) 4 个裁判删除最高最低取中间两位, 即 8.5, 8.5, 8.5, 8.5。
 (2) 同步 (SY) 5 个裁判删除最高最低取中间 3 位, 即 9.0, 9.0, 9.0, 9.0, 9.5。
 (3) 上述 5 位裁判分数相加, $8.5 + 8.5 + 9.0 + 9.0 + 9.0 = 44$ 。

(4) 分数乘以 $3/5$ (因为传统只有 3 个裁判分数), $44 \times 3/5 = 26.4$ 。

(5) 再乘以困难度, $26.4 \times 2 = 52.8$ 。

2.6 相对位置量数

相对位置量数有百分位数、十分位数、四分位数、中位数, 后 3 者是百分位数的特例。相对位置量数是箱线图的主要描述统计值。

2.6.1 百分位数

百分位数 (percentiles) 是将数据切成 100 份, 再计算其相对位置, 如果数据没有 100 个, 则按照近似切隔。切隔的方法有两种: 一种是按照数据的数目来切隔, 一种是按照数据的排序后的间距来切隔, 这个间距不是数据的实际值的差距, 例如, 第 5 个数据和第 6 个数据也许数值相等, 但是间距还是算 1。

百分位序或百分位秩 (percentrank) 是计算一个数值 x 在 n 个数据 x_1, x_2, \dots, x_n 中, 等于多少百分位数。例如: 一个学生高考、GRE 的分数, 在所有考生的分数中是多少百分位序。百分位序是由小到大的次序, “1 - 百分位序” 才是分数的前百分之几。

假设 x_1, x_2, \dots, x_n 等 n 个数据, 已由小到大排列。

(1) 百分位数: 如果数据分布 X , 输入 k , 求 P_k , 使 $P(X \leq P_k) = k/100$ 。

(2) 百分位序: 如果数据分布 X , 输入 x , 求 $P(X \leq x)$ 。

百分位数和百分位序是互为反函数。概率符号请见第 6, 7 章。

定义 假设一群数字数据有 x_1, x_2, \dots, x_n 等 n 个数据, 已按照大小, 由小到大排列, 即 $x_1 \leq x_2 \leq \dots \leq x_n$ 。第 k 个百分位数 (percentiles), 记作 P_k , 是有 $k\%$ 个观察值小于 P_k , 而有 $(100 - k)\%$ 个观察值大于 P_k 。(这个定义不完全精确, 因为加上 P_k , 超过 100%)

因此, 计算百分位数有许多公式, 以下介绍 3 种, 可以只选一种。

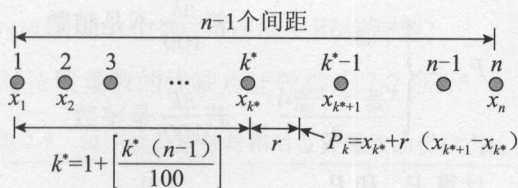
1. 以数据间距, 计算百分位数 P_k , Excel 公式 Percentile.INC (array, k), $0 \leq k \leq 1$

如果以数据的间距来计算百分位数 P_k , n 是数据的数目, 则:

(1) $k^* = [1 + \frac{nk}{100} - \frac{k}{100}]$ 是等于或小于 $1 + \frac{nk}{100} - \frac{k}{100}$ 的最大整数, 则

$$r = 1 + \frac{nk}{100} - \frac{k}{100} - k^*$$

$$(2) P_k = x_{k^*} + r(x_{k^*+1} - x_{k^*})。$$



例如：有 50 个数据，计算 P_{30} 。

$$k^* = [50 \times 0.3 + 1 - 0.3] = [15.7] = 15$$

$$r = 15.7 - 15 = 0.7$$

$$P_{30} = x_{15} + 0.7(x_{16} - x_{15})$$

(3) 百分位序：输入 x ，Excel 使用的公式 Percentrank.INC (array, x)。

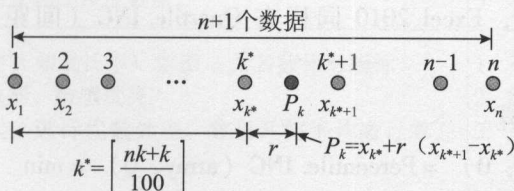
找出最大的 k ，使 $x_k \leq x$ ，令 $m = k + \frac{x - x_k}{x_{k+1} - x_k} - 1$ ，则 x 的百分位序 $p = m \times \frac{100}{(n-1)}$ 。

2. 以数据的个数来计算百分位数，Excel 公式 Percentile.EXC (array, k)， $0 \leq k \leq 1$

(1) $k^* = [\frac{nk+k}{100}]$ ，是等于或小于 $\frac{nk+k}{100}$ 的最大整数，则

$$r = \frac{nk+k}{100} - k^*$$

$$(2) P_k = x_{k^*} + r(x_{k^*+1} - x_{k^*})。$$



例如：有 50 个数据，计算 P_{30} 。

$$k^* = [\frac{nk+k}{100}] = [\frac{50 \times 30 + 30}{100}] = [15.3] = 15$$

$$r = 15.3 - 15 = 0.3$$

$$P_{30} = x_{15} + 0.3(x_{16} - x_{15})$$

(3) 百分位序：输入 x ，Excel 使用的公式 Percentrank.EXC (array, x)。

找出最大的 k ，使 $x_k \leq x$ ，令 $m = k + \frac{x - x_k}{x_{k+1} - x_k}$ ， x 的百分位序 $p = m \times \frac{100}{(n+1)}$ 。

3. 计算百分位数 P_k 的近似公式

$$P_k = \begin{cases} x_{[\frac{nk}{100}] + 1}, & \text{若 } \frac{nk}{100} \text{ 不是整数} \\ \frac{x_{\frac{nk}{100}} + x_{\frac{nk}{100} + 1}}{2}, & \text{若 } \frac{nk}{100} \text{ 是整数} \end{cases}$$

例如：有 50 个数据，计算 P_{30} 和 P_{85} 。

P_{30} ：由于 $\frac{nk}{100} = \frac{50 \times 30}{100} = 15$ ，是整数，所以 $P_{30} = \frac{x_{15} + x_{16}}{2} = x_{15} + 0.5(x_{16} - x_{15})$ 。

P_{85} ：由于 $\frac{nk}{100} = \frac{50 \times 85}{100} = 42.5$ ，不是整数，所以 $P_{85} = x_{43}$ 。

例题 2.19 学生的成绩百分位数和百分位序。（解答见网络资源）

2.6.2 四分位数

四分位数是将数据分成四等份，可以用在方盒图或箱线图中，表示数据集散的程度。下四分位数即第 25 个百分位数，中位数即第 50 个百分位数，上四分位数即第 75 个百分位数。

定义 假设一群数字数据有 x_1, x_2, \dots, x_n 等 n 个数据，已按照大小，由小到大排列，即 $x_1 \leq x_2 \leq \dots \leq x_n$ 。四分位数（quartiles），记作 Q_i ， $i=1, 2, 3$ ，是指至少有 $i/4$ 的观察值在 Q_i 之前，有 $(4-i)/4$ 的观察值在 Q_i 之后。

$$Q_1 = P_{25}, Q_2 = P_{50} = M_e, Q_3 = P_{75}$$

对于四分位数公式，Excel 2010 同样有 Quartile. INC（间距）和 Quartile. EXC（个数）。

（1）以数据间距计算。

$$\text{Quartile. INC}(\text{array}, 0) = \text{Percentile. INC}(\text{array}, 0) = \min$$

$$\text{Quartile. INC}(\text{array}, 1) = \text{Percentile. INC}(\text{array}, 0.25) = Q_1$$

$$\text{Quartile. INC}(\text{array}, 2) = \text{Percentile. INC}(\text{array}, 0.5) = Q_2 = M_e$$

$$\text{Quartile. INC}(\text{array}, 3) = \text{Percentile. INC}(\text{array}, 0.75) = Q_3$$

$$\text{Quartile. INC}(\text{array}, 4) = \text{Percentile. INC}(\text{array}, 1) = \max$$

（2）以数据个数计算

$$\text{Quartile. EXC}(\text{array}, 0) = \text{Percentile. EXC}(\text{array}, 0) = \min$$

$$\text{Quartile. EXC}(\text{array}, 1) = \text{Percentile. EXC}(\text{array}, 0.25) = Q_1$$

$$\text{Quartile. EXC}(\text{array}, 2) = \text{Percentile. EXC}(\text{array}, 0.5) = Q_2 = M_e$$

$$\text{Quartile. EXC}(\text{array}, 3) = \text{Percentile. EXC}(\text{array}, 0.75) = Q_3$$

Quartile. EXC (array, 4) = Percentile. EXC (array, 1) = max

例题 2.20 学生的成绩四分位数。（解答见网络资源）

集中趋势量数与相对位置量数的优缺点比较如表 2-7 所示。

表 2-7 集中趋势量数与相对位置量数的优缺点比较

代表值	优点	缺点
算术平均数	<div>1. 常识性代表值，数值重心</div> <div>2. 推断统计的主角</div> <div>3. 能列出公式，进行代数处理，由数组算术平均数计算总平均</div> <div>4. 代表所有数据，敏感度高，因为每个数据值都用来计算</div> <div>5. 数值与平均数之差的平方和为最小</div> <div>6. 数据经线性转换后，平均数可转换</div>	<div>1. 若有极端值，则代表性较差</div> <div>2. 若有偏态，则代表性较差（偏态表示有一边是长尾巴，可能有极端值）</div>
中位数	<div>1. 代表数据的中间值（制作箱线图）</div> <div>2. 能用于有极端值的数据</div> <div>3. 适用于有偏态的数据</div> <div>4. 能做无母数统计（适合顺序尺度数据）</div> <div>5. 数值与中位数差的绝对值之和为最小</div>	<div>1. 不能做代数处理</div> <div>2. 代表性较差，敏感度低</div> <div>3. 不能做有参数统计推论</div> <div>4. 数据要先排序</div>
众数	<div>1. 代表数量最多的数据</div> <div>2. 容易了解（较适合分类尺度数据）</div> <div>3. 能用于有极端值的数据</div> <div>4. 适用于有偏态的数据</div> <div>5. 适用于分类或顺序尺度数据</div>	<div>1. 可能不存在或不只一个</div> <div>2. 代表性差，敏感度低</div> <div>3. 不能做统计推论</div> <div>4. 数据不集中，则无意义</div> <div>5. 不适合连续型数据</div>
几何平均数	<div>1. 适合比率性（如成长率）数据（无名数相对指标）</div> <div>2. 代表所有数据，敏感度高</div> <div>3. 能列出公式，进行代数处理，数组几何平均数计算总平均</div>	<div>1. 不适合一般数据（绝对数）</div> <div>2. 数据不能有 0 或负数</div> <div>3. 只适合比率尺度数据</div> <div>4. 不能做推断统计</div>
调和平均数	<div>1. 适合有单位的数值数据（有名数相对指标） 例如：速率、物价数据</div> <div>2. 代表所有数据，敏感度高</div> <div>3. 能列出公式，进行代数处理</div>	<div>1. 不适合一般数据</div> <div>2. 数据不能有 0 或负数</div> <div>3. 只适合比率尺度数据</div> <div>4. 不能做推断统计</div>
截尾平均数	适用于有极端大小值的数据	不代表所有数据
四分位数	<div>1. 计算四分位距（制作箱线图）</div> <div>2. 作为数据的分界点</div>	同中位数
百分位数	作为数据的分界点	同四分位数

2.7 离差量数

离差量数 (measures of variability) 表示数据乖离变化的分布情况。离差量数可以让我们知道代表值的信赖度, 离差值越小, 代表值的代表性越大。离差值越大, 代表值的代表性越小, 可能是风险大, 可能是质量不稳定, 也可能是分布不平均。有人喜欢方差大追求风险刺激, 有人喜欢方差小希望稳定保守, 那是表现在个人生活或理财。在企业管理, 多数是要求离差值越小越好, 使产品质量一致或投资风险降低。

离差值有: 异众比率、极差、四分位差、方差、标准差、平均差、相对离差。离差值越大, 集中趋势量数 (平均指标代表值) 的代表性就越差; 离差值越小, 集中趋势量数 (平均指标代表值) 的代表性就越强。

2.7.1 异众比率

定类数据的离差值是异众比率 (variation rate), 非众数组的频数之和占总频数的比率称为异众比率 V_r ,

$$V_r = \frac{\sum f_i - f_{M_0}}{\sum f_i}$$

式中: $\sum f_i$ —— 变量值的总频数;

f_{M_0} —— 众数组的频数。

例题 2.21 例题 2.4 的异众比率。(解答见网络资源)

2.7.2 极差与四分位差

定序数据的离差值是全距和四分位差。全距是全部数据的范围, 即最大值和最小值之差。四分位差是中间一半数据的范围, 即下四分位数和上四分位数之差。

定义 假设一群数字数据有 x_1, x_2, \dots, x_n 等 n 个总体或样本数据, 已按照大小, 由小到大排列, 即 $x_1 \leq x_2 \leq \dots \leq x_n$ 。令下四分位数和上四分位数分别为 Q_1 和 Q_3 。则全距或称极差或全距 (range) R 和四分位差 (interquartile range) Q 分别为

$$R = x_n - x_1$$

$$Q = Q_3 - Q_1$$

抽样全距会受样本数目影响, 样本数目越大, 样本全距就会越大。如果数据呈正态钟

形分布, 则标准差大概是全距的 $1/4$ 。

例题 2.22 学生成绩的极差与四分位差。(解答见网络资源)

2.7.3 方差与标准差

定距数据的离差值是方差与标准差。

定义 假设一群数字数据有 x_1, x_2, \dots, x_n 等 n 个数据, 分以下两种情况来定义方差 (variance)。

(1) 这群数据是总体的全部数据, 则方差

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \frac{\sum_{i=1}^N x_i^2 - 2\mu \sum_{i=1}^N x_i + \sum_{i=1}^N \mu^2}{N} = \frac{\sum_{i=1}^N x_i^2}{N} - \mu^2 = \frac{\sum_{i=1}^N x_i^2}{N} - \frac{(\sum_{i=1}^N x_i)^2}{N^2}$$

(2) 这群数据是样本数据, 则方差

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} = \frac{\sum_{i=1}^n x_i^2}{n-1} - \frac{(\sum_{i=1}^n x_i)^2}{n(n-1)}$$

方差是各变量值与均值的离差平方的平均数。离差 = 变量值 - 均值。

方差 = Σ (变量值 - 均值)² / 自由度 = 离“差”平“方”的平“均”数 = 均方差

“方差”实际上应该是“均方差”, “方差”当作“均方差”的简称。

请注意, 为什么样本数据的方差公式的分母要用 $n-1$? 因为样本数据方差的分母若用 n , 则会低估总体方差 σ^2 以及自由度的问题 (请见第9章)。

平均平方差 (mean squared deviation) 是各数值与样本平均数之差的平方和的平均, 记作 MSD, 即

$$\text{MSD} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

定义 方差的正平方根, 称为标准差 (standard deviation)。

(1) 数据是总体的全部数据, 则标准差

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} = \sqrt{\frac{\sum_{i=1}^N x_i^2}{N} - \mu^2}$$

(2) 数据是样本数据, 则标准差

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}}$$

由于方差的单位是原来数据的单位（例如：cm、kg）的平方，所以标准差与平均数有相同的单位，就可以做加减运算，例如： $\mu \pm 2\sigma$ 。

标准差与方差的特点如下。

- (1) 配合算术平均的离差值，方差越小，平均数的代表性越强。
- (2) 推断统计的主要参数。
- (3) 能将数据标准化（分子分母单位相同），变成平均数为 0，方差为 1 的无单位数值：

$z_i = \frac{x_i - \mu}{\sigma}$ ，称为 Z 分数或标准分数，即 x_i 和平均数 μ 的距离是 $|z_i|$ 个标准差 σ ，

$|x_i - \mu| = |z_i| \sigma$ 。如果是样本数据，则用 \bar{x} 代替 μ 。Z 分数越大表示 x_i 数据排序越高，Z 分数越小（负数）表示 x_i 数据排序越低。

例题 2.23 Z 分数。（解答见网络资源）

- (4) 能进行代数处理，计算数组标准差的总和。例如：“相同”总体数据分成 m 组，每组 n_i 数据，总数 $N = \sum_{i=1}^m n_i$ 。若每组平均数为 μ_i ，每组方差为 σ_i^2 ，则总平均数 μ ，总方差 σ^2 分别为

$$\mu = \frac{\sum_{i=1}^m n_i \mu_i}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^m n_i [\sigma_i^2 + (\mu_i - \mu)^2]}{N}$$

m 组样本数据，来自“相同”总体（相同平均数及方差），每组 n_i 数据，总数 $n = \sum_{i=1}^m n_i$ 。若每组平均数为 \bar{x}_i ，每组方差为 s_i^2 ，则总平均数 \bar{x} ，总方差 s^2 分别为

$$\bar{x} = \frac{\sum_{i=1}^m n_i \bar{x}_i}{n}$$

$$s^2 = \frac{\sum_{i=1}^m [(n_i - 1)s_i^2 + n_i (\bar{x}_i - \bar{x})^2]}{n - 1}$$

如果两（ m ）组独立抽样的样本数据，来自两（ m ）个“不同”总体（平均数可能不同），但两（ m ）总体“方差相同”（11.3 节），则两（ m ）组样本数据，可合并计算其共同的方差。若每组样本容量为 n_i ，每组样本方差为 s_i^2 ，则合并方差 s_{pool}^2 （或记作 s_p^2 ）为

(请见 11.4 节及 12.4 节的 MS_E 公式)

$$s_{\text{pool}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

$$s_{\text{pool}}^2 = \frac{\sum_{i=1}^m (n_i - 1)s_i^2}{n - m}$$

(5) 代表所有数据, 敏感度高, 因为每个数据值都用来计算。

(6) 可以代入切比雪夫不等式。

(7) 方差为各数值与任何常数之差的平均平方和最小。

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \leq \frac{\sum (x_i - r)^2}{N}, \quad \forall r \in R$$

(8) 若一组数据 (总体或抽样) x_i , 经过线性转换: $y_i = a + bx_i$, a 和 b 是实数,

则 x_i 的方差 σ_x^2 与 y_i 的方差 σ_y^2 的关系为: $\sigma_y^2 = b^2 \sigma_x^2$ 。

(9) 如果数据呈正态钟形分布, 则标准差大概是全距的 $1/4$: $s = R/4$ 。

(10) 如果 $\sigma = 0$ 或 $s = 0$, 则所有数据都为相同值。

例题 2.24 学生成绩的方差和标准差。(解答见网络资源)

2.7.4 平均差

每个观测值和代表值之差, 称为离差 (deviation)。以平均数为中心的离差称为离均差 (deviation about the mean), 以中位数为中心的离差称为离中差 (deviation about the median)。离均差有正数也有负数, 离均差全部加起来等于 0, 即

$$\sum_{i=1}^n (x_i - \mu) = \sum_{i=1}^n (x_i - \bar{x}) = 0$$

所以, 计算离差值的测定, 除了平方和, 就是取绝对值, 但是绝对值不能做代数运算, 如上述方差特点的第 4、8 点, 也不能做统计推论。

定义 假设一群数字数据有 x_1, x_2, \dots, x_n 等 n 个数据, 平均差 (mean deviation) 是各数值与平均数 (或中位数) 之差的绝对值的平均。

(1) 以平均数为中心的平均差或称为平均离均差 MD_μ , 其计算公式为

$$MD_\mu = \sum_{i=1}^n |x_i - \mu| / n$$

(2) 样本数据平均绝对差 (mean absolute deviation, MAD), 其计算公式为

$$MAD = \sum_{i=1}^n |x_i - \bar{x}| / n$$

(3) 以中位数为中心的平均差或称为平均离中差 MD_{M_e} ，其计算公式为

$$MD_{M_e} = \sum_{i=1}^n |x_i - M_e| / n$$

$$MD_{M_e} = \sum_{i=1}^n |x_i - M_e| / n \leq \sum_{i=1}^n |x_i - r| / n, \quad \forall r \in R$$

虽然以中位数计算的平均差 MD_{M_e} 为最小，但是通常使用平均离均差 MD_{μ} （总体数据）或平均绝对差（样本数据）MAD。

例题 2.25 学生成绩的平均差。（解答见网络资源）

2.7.5 相对离差

上述全距、四分位差、标准差、平均差等离差衡量都与原数据有相同的单位，所以称为“绝对离差”。方差的单位为原数据单位的平方。如果数据的衡量尺度、单位、平均数都相同，则可以用绝对离差加以比较。

如果数据的衡量尺度不同，或单位不同，或平均数不同，则要用相对离差来比较。相对离差是绝对离差除以代表值（平均数），是无单位的数值。使用相对离差时，原始数据最好是正数，否则平均数接近 0 或为负数，则相对离差无意义。相对离差有：变异系数、平均差系数、全距系数、四分位差系数。

(1) 变异系数（variation coefficient）VC，其计算公式为

$$VC = \frac{\sigma}{\mu} \quad \text{或} \quad VC = \frac{s}{\bar{x}}$$

使用变异系数，要注意每个数据，必须为正数，或几乎都是正数。

(2) 平均差系数（mean deviation coefficient）MC，其计算公式为

$$MC = \frac{MD_{\mu}}{\mu} \quad \text{或} \quad MC = \frac{MD_M}{M_e}$$

(3) 全距系数（range coefficient）RC，其计算公式为

$$RC = \frac{x_{\max} - x_{\min}}{x_{\max} + x_{\min}} = \frac{x_n - x_1}{x_n + x_1}$$

(4) 四分距系数（interquartile range coefficient）QC，其计算公式为

$$QC = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

例题 2.26 学生成绩的相对离差。（解答见网络资源）

2.8 形态量数

2.8.1 三阶距与四阶距

三阶距与四阶距是计算偏度系数和峰度系数的必要衡量值。

定义 假设一群数字数据有 x_1, x_2, \dots, x_n 等 n 个数据, 定义三阶距 (third moment):

(1) 总体数据, 则三阶原点距 M'_3 , 三阶中心距 M_3 , 分别为

$$M'_3 = \frac{\sum_{i=1}^N x_i^3}{N}$$

$$M_3 = \frac{\sum_{i=1}^N (x_i - \mu)^3}{N} = \frac{\sum x_i^3 - 3\mu \sum x_i^2 + 3\mu^2 \sum x_i - N\mu^3}{N}$$

(2) 样本数据, 则三阶原点距 m'_3 , 三阶中心距 m_3 , 分别为

$$m'_3 = \frac{\sum_{i=1}^n x_i^3}{n-1}$$

$$m_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n-1} = \frac{\sum x_i^3 - 3\bar{x} \sum x_i^2 + 3\bar{x}^2 \sum x_i - n\bar{x}^3}{n-1}$$

定义 假设数据 x_1, x_2, \dots, x_n , 定义四阶距 (fourth moment):

(1) 总体数据, 四阶原点距 $M'_4 = \frac{\sum_{i=1}^N x_i^4}{N}$, 四阶中心距 $M_4 = \frac{\sum_{i=1}^N (x_i - \mu)^4}{N}$ 。

(2) 样本数据, 四阶原点距 $m'_4 = \frac{\sum_{i=1}^n x_i^4}{n-1}$, 四阶中心距 $m_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n-1}$ 。

一阶原点距 $M'_1 = \mu$, 一阶中心距 $M_1 = 0$ 。

二阶原点距 $M'_2 = \sum x^2/N$, 二阶中心距 $M_2 = \sigma^2$ 。

例题 2.27 学生成绩的三阶距与四阶距。(解答见网络资源)

2.8.2 偏度

偏度 (skewness) 表示数据分布的“对称”情况，有 3 种类型：对称型 (symmetrical)，右偏型 (skewed right)，左偏型 (skewed left)。对称型是平均数、中位数、众数都在相同点。右偏型：高峰（众数）在左，平均数在右；如果是单峰型态，则众数小于中位数，中位数小于平均数。左偏型：高峰（众数）在右，平均数在左；如果是单峰型态，则平均数小于中位数，中位数小于众数。

由于左偏型是左边的尾巴较长（较斜），右边的概率较高；右偏型是右边的尾巴较长（较斜），左边的概率较高。所以左偏型较正确的翻译应该是“左斜型”；右偏型较正确的翻译应该是“右斜型”。

各种类型偏度的形状如图 2-17 所示。

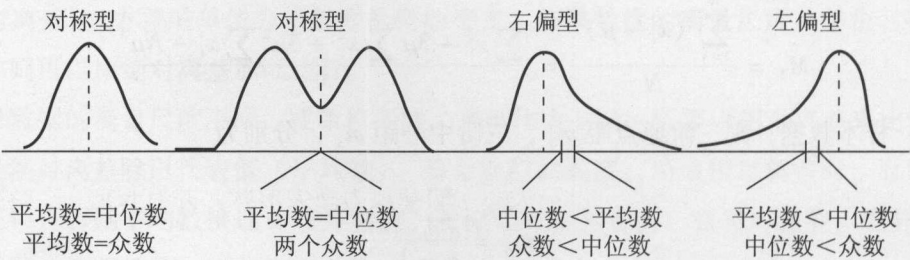


图 2-17 偏度形状

定义 假设一群数字数据有 x_1, x_2, \dots, x_n 等 n 个数据，若偏度系数等于 0，则为对称型；若偏度系数大于 0，则为右偏型， $M_0 < M_d < \mu$ ；若偏度系数小于 0，则为左偏型， $\mu < M_d < M_0$ 。

以下几个方法来定义偏度系数 SK。

1) 皮尔生 (Pearson) 偏度指数。

(1) 总体数据，SK 的计算公式为

$$SK = \frac{3(\mu - M_d)}{\sigma}$$

(2) 样本数据，SK 的计算公式为

$$SK = \frac{3(\bar{x} - M_d)}{s}$$

2) 利用三阶距，计算偏度系数。

(1) 总体数据，SK 的计算公式为

$$SK = \frac{M_3}{\sigma^3}$$

(2) 样本数据, SK 的计算公式为

$$SK = \frac{m_3}{s^3}$$

3) 利用 Excel 的公式, 计算偏度系数。

(1) 总体数据, SK 的计算公式为

$$SK = \frac{nM_3}{(n-2)\sigma^3}$$

(2) 样本数据, SK 的计算公式为

$$SK = \frac{nm_3}{(n-2)s^3}$$

例题 2.28 学生成绩的偏度。(解答见网络资源)

2.8.3 峰度

峰度 (kurtosis) 表示数据分布的情况, 有 3 种类型: 正态峰型 (mesokurtic), 尖峰型 (leptokurtic), 平峰型 (platykurtic)。如果在相同的标准差, 则正态峰型像正态曲线; 平峰型的峰顶为扁平形状, 并有短尾巴; 尖峰型的峰顶为高尖形状, 并有长或厚的尾巴, 如图 2-18 所示。

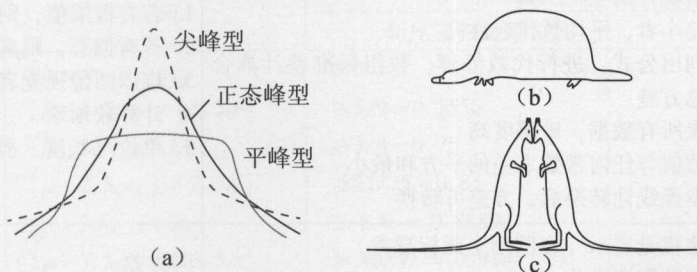


图 2-18 有相同的标准差的峰度形状: 正态峰、平峰、尖峰

(a) 有相同的标准差的 3 种峰; (b) 平峰型; (c) 尖峰型

长方形的均匀分布为平峰型的代表, 四阶距峰度系数为 1.8; 指数分布为尖峰型的代表, 四阶距峰度系数为 9; 正态分布是正态峰型, 四阶距峰度系数为 3。

定义 假设数据 x_1, x_2, \dots, x_n , 定义峰度系数 K :

1) 利用四阶距, 计算峰度系数。

(1) 总体数据, K 的计算公式为

$$K = \frac{M_4}{\sigma^4}$$

(2) 样本数据， K 的计算公式为

$$K = \frac{m_4}{s^4}$$

若峰度系数等于 3，则为正态峰型；若峰度系数大于 3，则为尖峰型；若峰度系数小于 3，则为平峰型。

均匀分布的峰度是 1.8，是标准的平峰型；指数分布的峰度是 9，是标准的尖峰型。

2) Excel 的公式，计算峰度系数。

样本数据， K 的计算公式为

$$K = \frac{n(n+1)m_4}{(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}, \quad s \neq 0, n \geq 4$$

若峰度系数等于 0，则为正态峰型；若峰度系数大于 0，则为尖峰型；若峰度系数小于 0，则为平峰型。

例题 2.29 学生成绩的峰度。

离数指标与形态指标的优缺点比较如表 2-8 所示。

表 2-8 离差指标与形态指标的优缺点比较

代表值	优点	缺点
方差	1. 配合算术平均数的离差值 2. 推断统计的主要工具 4. 方差小者，平均数代表性高 5. 能列出公式，进行代数处理，数组标准差计算合并总方差 6. 代表所有数据，敏感度高 7. 各数值与任何常数之差的平方和最小 8. 数据经线性转换后，方差可转换	1. 若有极端值，则离差值不好 2. 若有偏态，则离差值不好 3. 难以图像视觉表示 4. 计算较烦琐 5. 单位（长度，质量）影响较大
标准差	1. 图像视觉表示（在图形上画标准差） 2. 能将数据标准化	同方差
异众比率	1. 适合定类尺度数据的离差衡量 2. 也可以用在定序或定距尺度数据	如果连续型数据没有众数或众数的频数很小，则虽然数据很集中，但异众比率仍很大
全距	1. 计算很简单，容易了解 2. 可用图像表示（箱线图） 3. 如果数据呈常态钟型分布，则标准差大概是全距的 1/4 4. 在质量管理图中，代替标准差的估计	1. 若有极端值，则结果差异大 2. 抽样全距受样本数影响 3. 没有代表值做配合 4. 中间值改变，敏感度低，无法知道中间数据的差异
四分位距	1. 配合中位数的离差值 2. 可用图像表示（箱线图） 3. 适合较偏态的资料	1. 计算较全距麻烦 2. 两端值改变，敏感度低 3. 数据数目少时，没什么意义

续表

代表值	优点	缺点
平均差	1. 配合算术平均数或中位数 2. 代表所有数据，敏感度高 3. 以中位数计算的平均差为最小 4. 较标准差更适合有极端值的情况	1. 不能做代数处理 2. 不能用于推断统计
相对离差	1. 单位不同之数据的差异比较 2. 单位相同但平均数不同之数据的差异比较	1. 不能用于推断统计 2. 平均数不能为 0 或负数
偏度系数	1. 单峰分配表示偏态形状 2. 判别数据是否对称（常态）	1. 不适合多峰分配 2. 小样本不准确
峰度系数	1. 单峰分配表示峰态形状 2. 判别数据是否为常态	1. 不适合多峰分配 2. 小样本不准确

2.8.4 切比雪夫定理

切比雪夫定理（Chebychev’s theorem）也可说明数据分布的一个特性，适合所有的母体数据或样本数据，也不限定数据分配的类型。

定理 假设一群数字数据有 x_1, x_2, \cdots, x_n 等 n 个数据，其平均数和标准差分别为 μ 和 σ ，则该数据在 $(\mu - k\sigma, \mu + k\sigma)$ 区间的概率等于 $1 - (1/k^2)$ ， $k > 1$ ，如表 2-9 所示。

表 2-9 切比雪夫定理和经验法则（正态分布概率）

k	区间	观测值落在该区间的概率	
		切比雪夫定理	正态分布概率
1.0	$[\mu - \sigma, \mu + \sigma]$	≥ 0	0.6826
1.5	$[\mu - 1.5\sigma, \mu + 1.5\sigma]$	$\geq 5/9 = 0.5556$	0.7062
2.0	$[\mu - 2\sigma, \mu + 2\sigma]$	$\geq 3/4 = 0.75$	0.9544
2.5	$[\mu - 2.5\sigma, \mu + 2.5\sigma]$	$\geq 21/25 = 0.84$	0.9876
3.0	$[\mu - 3\sigma, \mu + 3\sigma]$	$\geq 8/9 = 0.8889$	0.9974
3.5	$[\mu - 3.5\sigma, \mu + 3.5\sigma]$	$\geq 45/49 = 0.9184$	0.9978

例题 2.30 切比雪夫定理。（解答见网络资源）

2.9 中文统计应用

2.9.1 描述统计—原始数据（例题 2.2）

在 Excel 中，选择“加载项”→“中文统计”→“描述统计”→“原始数据”命令。在弹出的“描述统计—原始数据”对话框中执行下列操作。

（1）输入区域：输入 A1：A30，或用鼠标选取，在 A1 按住左键，往下拉到 A30。

- (2) 标志位于第一行：不要勾选。
- (3) 输出范围：输入单元格位置 C1 或选新工作表（可以不输入名称，有预设名称）。
- (4) 计算数值可用默认值或修改。
- (5) 按“确定”。操作如图 2-19 所示。

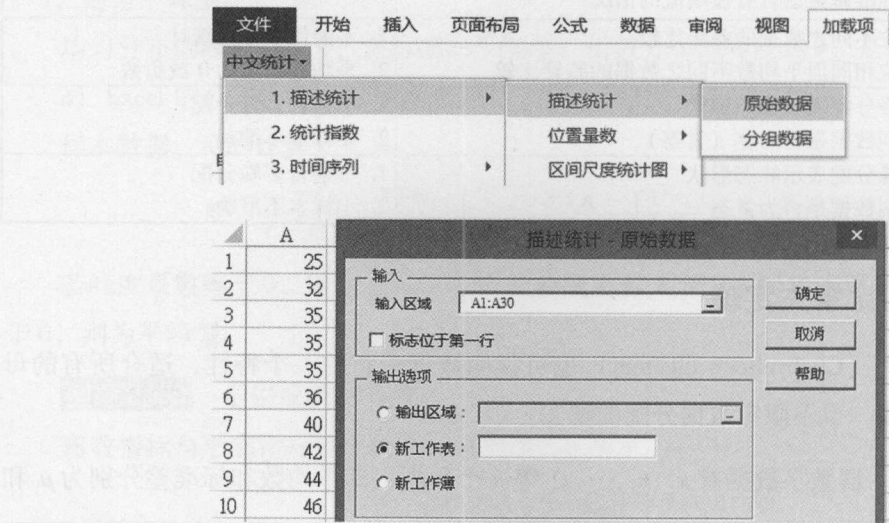


图 2-19 操作图示

“描述统计”的结果如图 2-20 所示。

	A	B	C	D	E	F	G
1	描述统计(原始数据):						
2	数据数目	30	总和	1800	最小值	25	最大值
3							
4	集中趋势量数						
5	算术均值	60				中位数	50
6	几何均值	55.82600585		调和均值	60.8	众数	35
7	调和均值	51.85599351		5%截尾均值	60		
8							
9	离散量数						
10		样本数据	总体数据	全距	70	平均离均差	20.2
11	方差	515.5862069	498.4	总体数据变异系数	0.372081234	平均离中差	19.2
12	标准差	22.70652344	22.32487402	样本数据变异系数	0.378442057	平均绝对差	20.2
13							
14	相对位置量数 - 四分位数 计算问题			四分位差 计算个数			
15	下四分位数	42.5		下四分位数	41.5		
16	中位数	50		中位数	50		
17	上四分位数	85.75		上四分位数	86.25		
18	四分位差	43.25		四分位差	44.75		
19							
20	相对位置量数 - 百分位数和百分位排序(百分位阶)						
21	p (%)	第 p 百分位数		百分位数 x			百分位排序(百分位阶 %)
22		计算问题	计算个数		计算问题	计算个数	
23	95	93.1	94.45	60	59.48	58.87096774	
24							
25	偏度与峰度系数						
26		样本数据			总体数据		
27		Excel公式	三级动差偏度系数	皮尔生偏度系数	SPSS公式	三级动差偏度系数	皮尔生偏度系数
28	偏度系数	0.304606805	0.284299685	1.32120622	0.309814137	0.289159861	1.343792577
29							
30		样本数据		总体数据			
31		Excel公式	四级动差偏度系数	四级动差偏度系数			
32	峰度系数	-1.465556981	1.521547228	1.574014374			

图 2-20 “描述统计”的结果

2.9.2 直方图（例题 2.2）

执行“直方图”的操作示意图和结果如图 2-21 和图 2-22 所示。

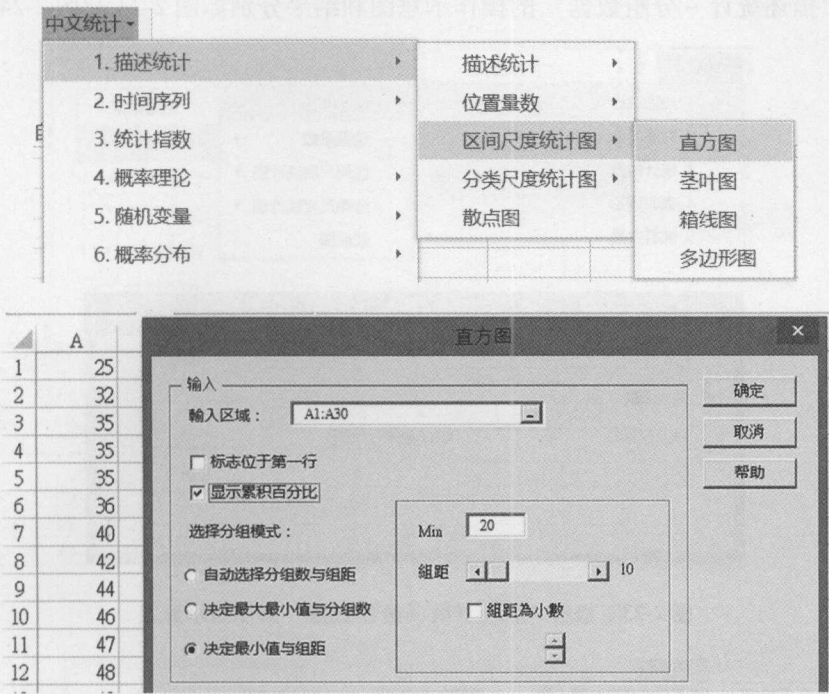


图 2-21 执行“直方图”的操作示意图

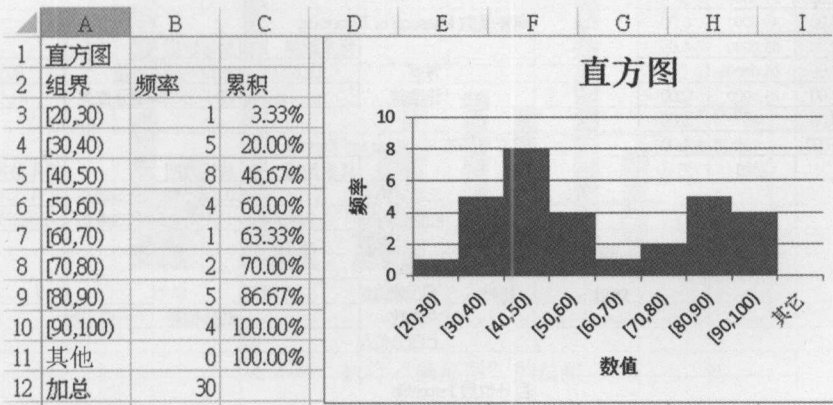


图 2-22 执行“直方图”的结果

2.9.3 描述统计 – 分组数据（例题 2.2 直方图的分组）

分组数据的描述统计，其计算方法及公式，请下载补充教材。

执行“描述统计 – 分组数据”的操作示意图和结果分别如图 2-23 和图 2-24 所示。

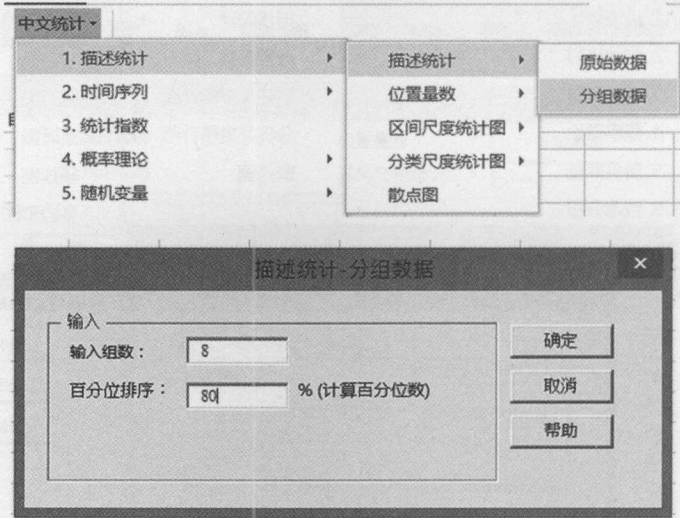


图 2-23 执行“描述统计 – 分组数据”的操作示意图

	A	B	C	D	E	F	G	H	I	J	K	L
1	描述统计（分组数据）				请在蓝色单元格输入数据：		组数：	8	红色是计算结果			
2												
3	组间		重新计算			集中趋势量数 Measures of Central tendency						
4	下组界	上组界	组中点 xI	频数 fI						金氏众数	86.6666667	
5	20.00	30.00	25.0000	1.00		均值Mean	60	中位数Median	52.5	苏伯众数	87.5	
6	30.00	40.00	35.0000	5.00								
7	40.00	50.00	45.0000	8.00		离散量数 Measures of Dispersion						
8	50.00	60.00	55.0000	4.00				样本数据	总体数据			
9	60.00	70.00	65.0000	1.00		方差	508.6206897	491.6666667	极差	80.00		
10	70.00	80.00	75.0000	2.00		标准差	22.55262046	22.1735578	四分位差	41.125		
11	80.00	90.00	85	5.00								
12	90.00	100.00	95	4.00		偏态与峰态 Skewness and Kurtosis						
13			总和	30.00				样本数据	总体数据		样本数据	总体数据
14						三级差偏	0.32466444	0.33021466	SPSS偏态	0.34785476	0.35360142	
15						四级差偏	1.586710141	1.64349325	SPSS峰态	-1.3829359	-1.315544	
16												
17	四分位数 Quartiles											
18						下四分位数	41.875					
19						中位数	52.5	四分位差	41.125			
20						上四分位数	83					
21												
22	百分位数 Percentile											
23						x	第 x 百分位数					
24						80.00	86					

图 2-24 执行“描述统计 – 分组数据”的结果

2.9.4 条形图 (例题 2.4)

执行“条形图”的操作示意图和结果分别如图 2-25 和图 2-26 所示。

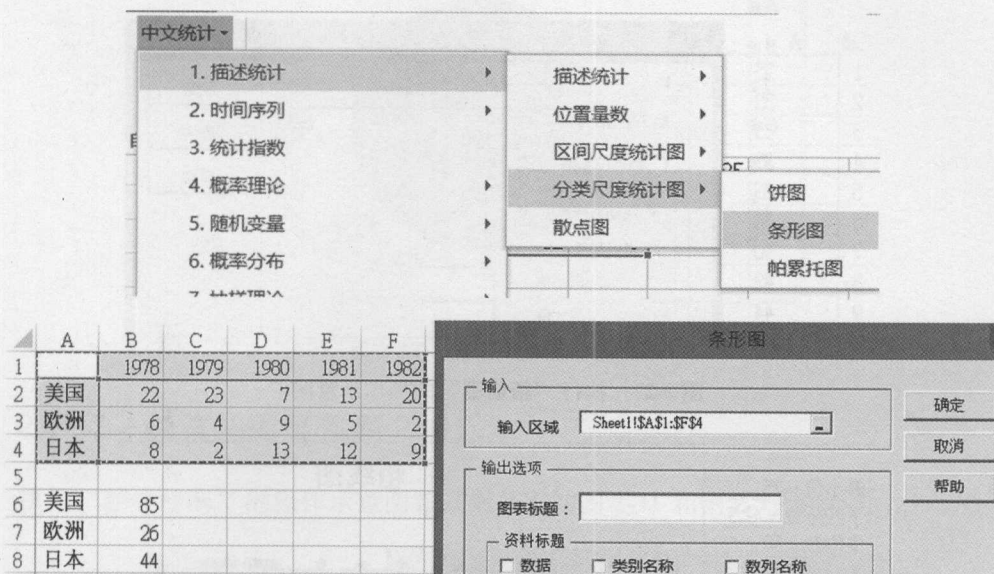


图 2-25 执行“条形图”的操作示意图

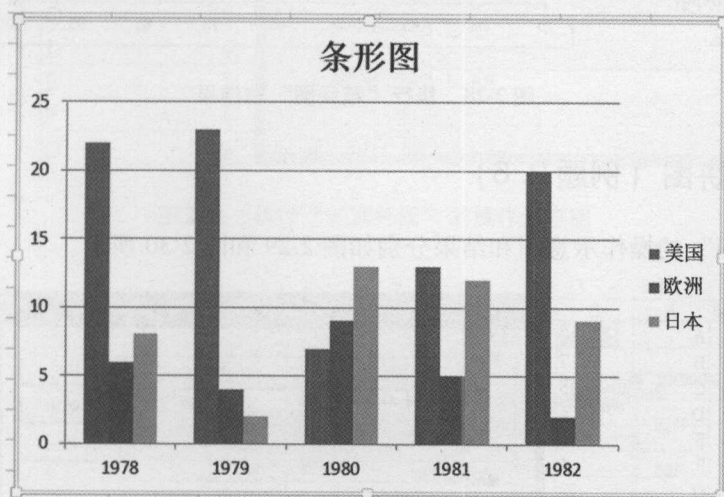


图 2-26 执行“条形图”的结果

2.9.5 箱线图（例题 2.5）

执行“箱线图”的操作示意图和结果分别如图 2-27 和图 2-28 所示。

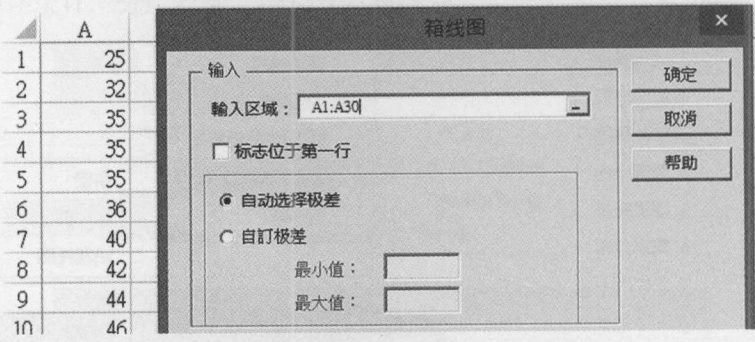


图 2-27 执行“箱线图”的操作示意图

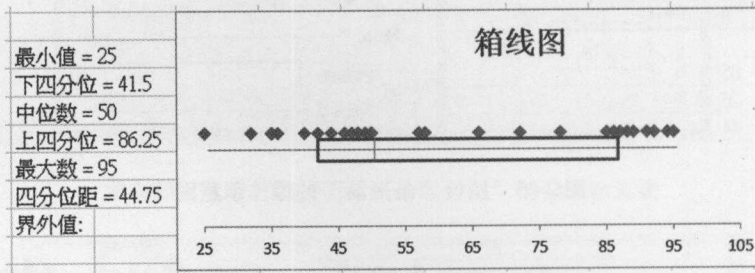


图 2-28 执行“箱线图”的结果

2.9.6 饼图（例题 2.6）

执行“饼图”的操作示意图和结果分别如图 2-29 和图 2-30 所示。

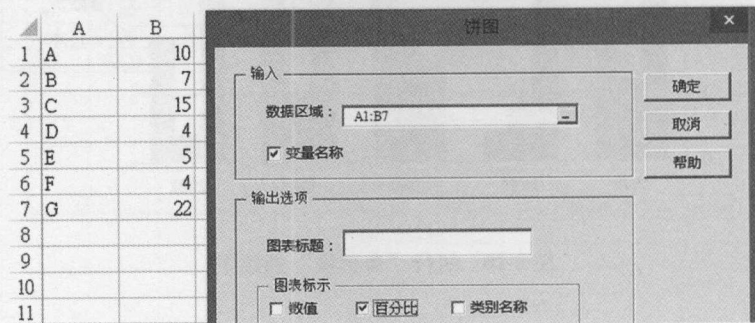


图 2-29 执行“饼图”的操作示意图

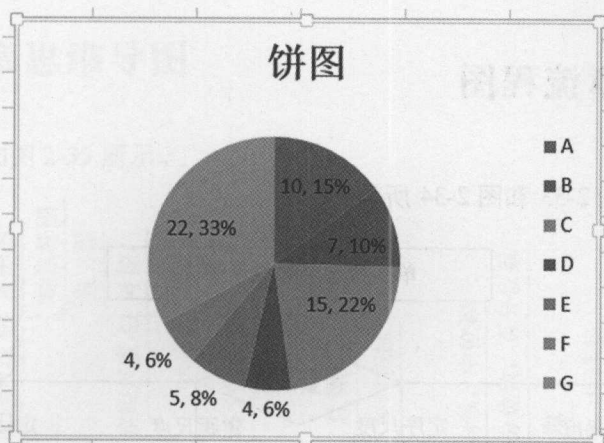


图 2-30 执行“饼图”的结果

2.9.7 帕累托图（例题 2.6）

执行“帕累托图”的操作示意图和结果分别如图 2-31 和图 2-32 所示。

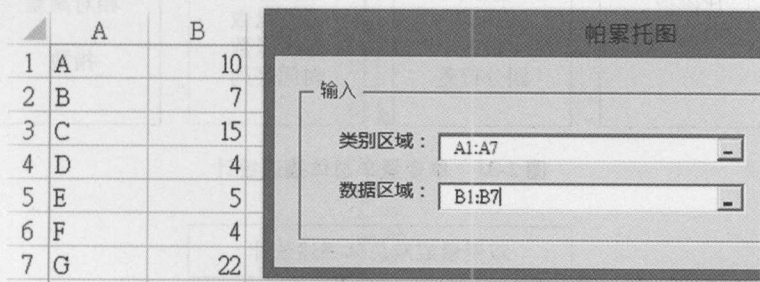


图 2-31 执行“帕累托图”的操作示意图

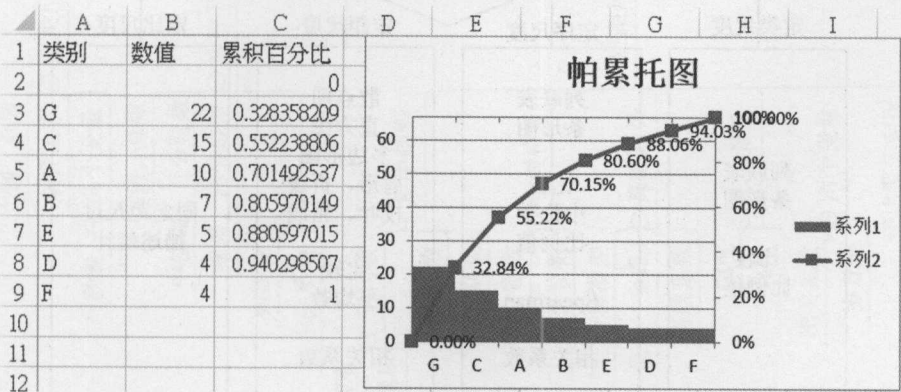


图 2-32 执行“帕累托图”的结果

2.10 本章流程图

本章流程图如图 2-33 和图 2-34 所示。

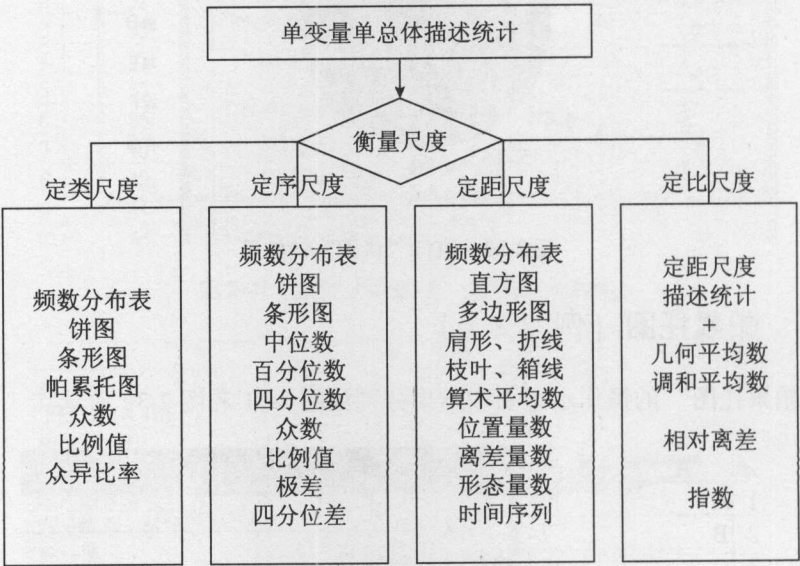


图 2-33 单变量单总体描述统计



图 2-34 双变量或双总体描述统计



2.11 本章思维导图

本章思维导图如图 2-35 所示。

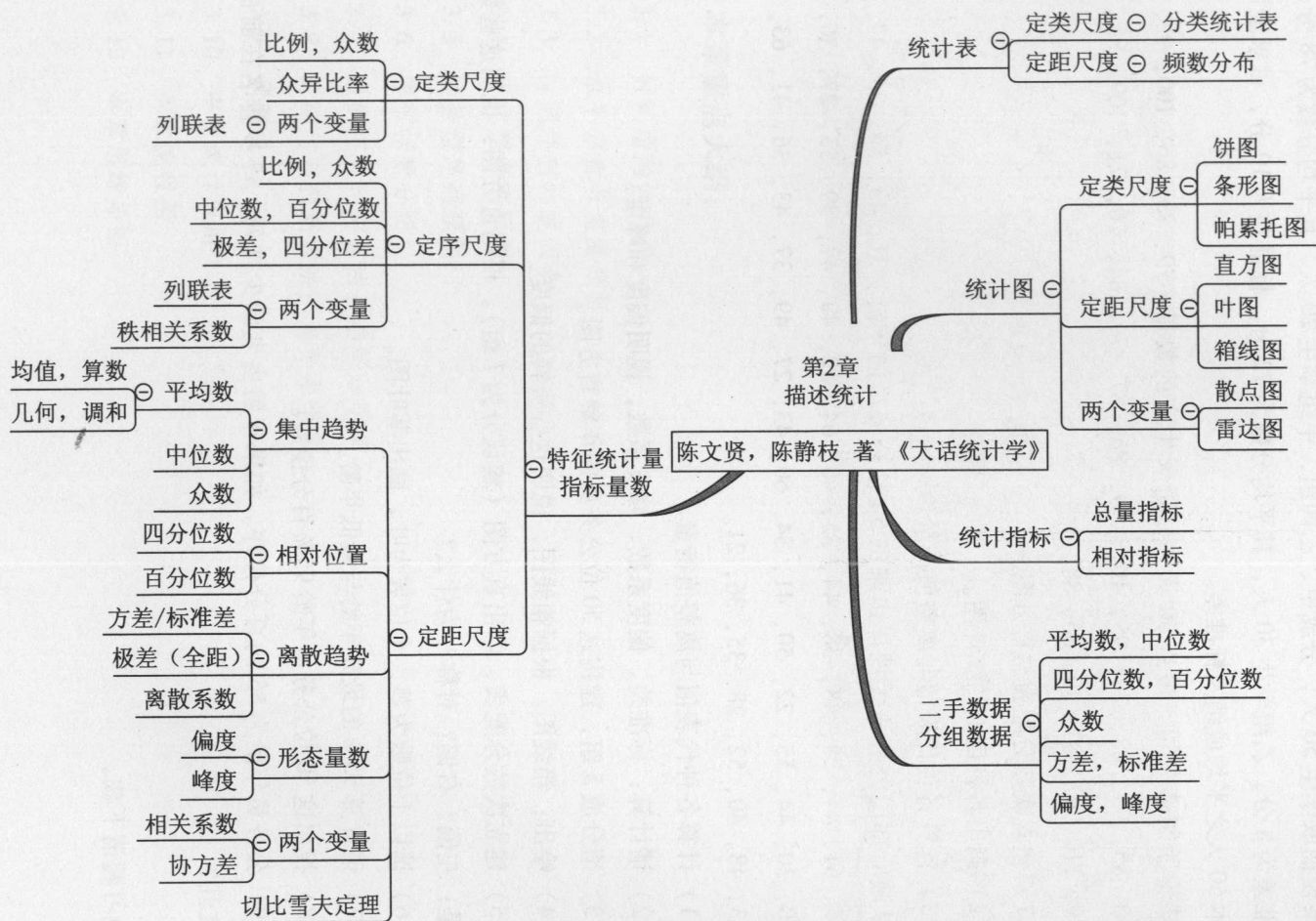


图2-35 第2章的思维导图

习题

1. 某校抽取学生 50 人，分成甲、乙两组，甲组学生 20 人，其平均分数为 78 分，标准差为 8 分；乙组学生 30 人，其平均分数为 72 分，标准差为 10 分，试求全部样本 50 人之平均成绩及标准差。
2. 某轮胎制造公司，最近 30 天中，每天生产的数量：79, 93, 86, 100, 92, 88, 80, 85, 93, 88, 78, 95, 101, 99, 86, 87, 79, 84, 76, 71, 109, 85, 79, 89, 110, 97, 93, 79, 86, 88。
 - (1) 建立次数分配表与直方图。
 - (2) 画出各种可能的统计图。
 - (3) 计算各种代表值与离差值衡量。
3. 从政府部门公务机关抽样出来的 50 位公务员的年龄：31, 43, 56, 23, 49, 42, 33, 61, 44, 28, 48, 38, 44, 35, 40, 64, 52, 42, 47, 39, 53, 27, 36, 35, 20, 30, 44, 55, 22, 50, 41, 34, 60, 43, 27, 49, 37, 43, 36, 41, 63, 51, 43, 48, 40, 52, 28, 35, 36, 21。
 - (1) 计算各种代表值与离差值衡量。
 - (2) 请计算：标准差、偏度系数、峰度系数，说明偏度和峰度。
 - (3) 请分成 5 组，建构这 50 位公务员的年龄直方图。
 - (4) 绘出：箱线图。根据箱线图，说明这些资料的偏度。
 - (5) 建立次数分配表，画出直方图（建议分为 7 组）。根据直方图，说明这些数据是：左偏？右偏？对称？为什么？
 - (6) 请以十位数为茎、个位数为叶，画出茎叶图。
 - (7) 请计算 95% 的最高年龄与最低年龄。
 - (8) 找出这 50 位公务员年龄 95% 百分位数。
 - (9) 公务员年龄为 55 的百分位序。如果年龄由老到少排列，55 岁排名在前百分之几？

其他习题请下载。



第3章

时间序列

人无远虑，必有近忧。

——《论语·卫灵公》

数往者顺，知来者逆，是故易逆数也。

——《易传·说卦》

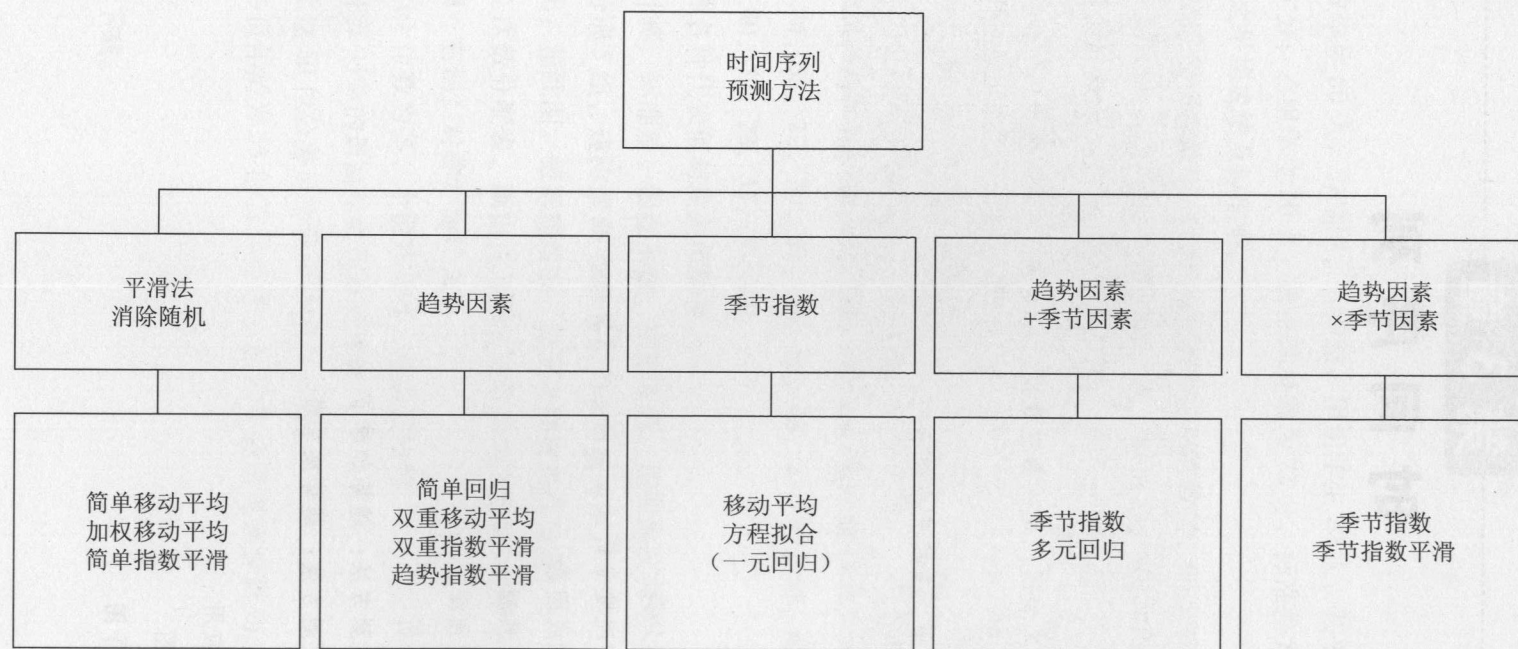
千金难买早知道，万般无奈想不到。

——俗语



本章重点大纲：

- 3.1 时间序列的分类
- 3.2 水平分析与速度分析
- 3.3 时间序列构成因素
- 3.4 平稳型序列预测
- 3.5 趋势型序列预测
- 3.6 季节指数分析
- 3.7 时间序列预测方法：趋势加季节
- 3.8 时间序列预测方法：趋势乘季节
- 3.9 预测误差
- 3.10 中文统计应用
- 3.11 本章流程图
- 3.12 本章思维导图



本章概念图

3.1 时间序列的分类

预测的主要目的是预估未来的现象，减少不确定的风险，作为计划与决策的基础。因此，预测方法越可靠，则计划结果越佳。

时间序列（time series）是过去时间的观察值（或称实际值）的数据。时间序列是一个变量在连续相等间隔（如每日，每周，每月）的时间内的观察记录。时间序列根据观察值的形式，比照指标的分类：总量指标、相对指标、平均指标，时间序列的分类可以分为：绝对数时间序列、相对数时间序列、平均数时间序列。

3.1.1 绝对数时间序列

绝对数指标按时间顺序排列的数列，可分为时期序列和时点序列，如图 3-1 所示。

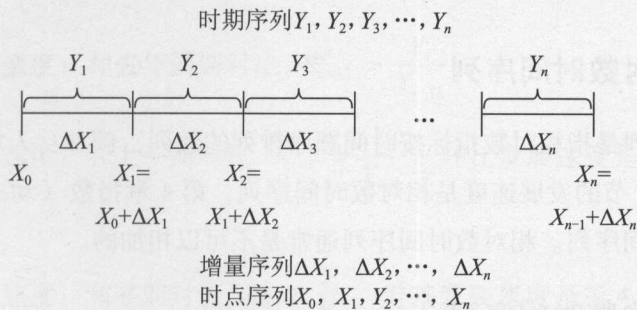


图 3-1 时期序列与时点序列

1. 时期序列

时期序列是反映在某一段时间的“总量”或结果，是累积这段时期的活动总量，例如：年国内生产总值是累积一年（时期）的生产（活动）总值。时期序列的观察值是可以相加的，而成为更长的时期。时期序列的观察值应该是时期越长，观察值越大，例如：“年生产量”的时间序列数值，一定大于“季生产量”的时间序列数值。但是，“年盈余总获利”的数值，不一定大于“季盈余总获利”的数值。因为，有可能某些季的盈余是亏损的（负数）。

2. 时点序列

时点序列是反映在某一个时间点的“水平”，例如：年底总人口数。股价 K 线图是时点序列，其日线（每日股价）包括当日的：开盘价、收盘价、最高价、最低价，如图 3-2 所示。

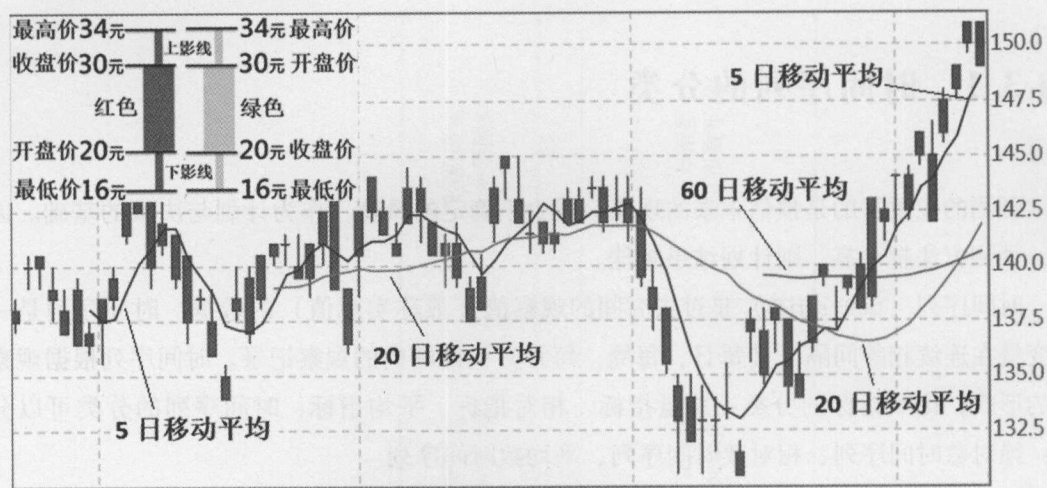


图 3-2 股价 K 线图和移动平均线

时点序列的观察值通常是不可以相加的，只有股票价格的移动平均线，是先相加再平均。

3.1.2 相对数时间序列

相对数时间序列是指相对数指标按时间顺序排列的数列，例如：人均生产总量是相对数时间序列，3.2.1 节的发展速度是相对数时间序列，第 4 章指数（动态相对指针）的时间数列是相对数时间序列。相对数时间序列通常是不可以相加的。

3.1.3 平均数时间序列

平均数时间序列是指平均数指标按时间顺序排列的数列，例如：平均工资序列。

3.2 水平分析与速度分析

3.2.1 水平分析

水平分析主要是计算平均发展水平，适用于绝对数、相对数和平均数序列，如图 3-3 所示。

3.2.2 速度分析

时间序列速度分析适用于“时期序列”的观察值数据 $Y_1, Y_2, Y_3, \dots, Y_n$ 。

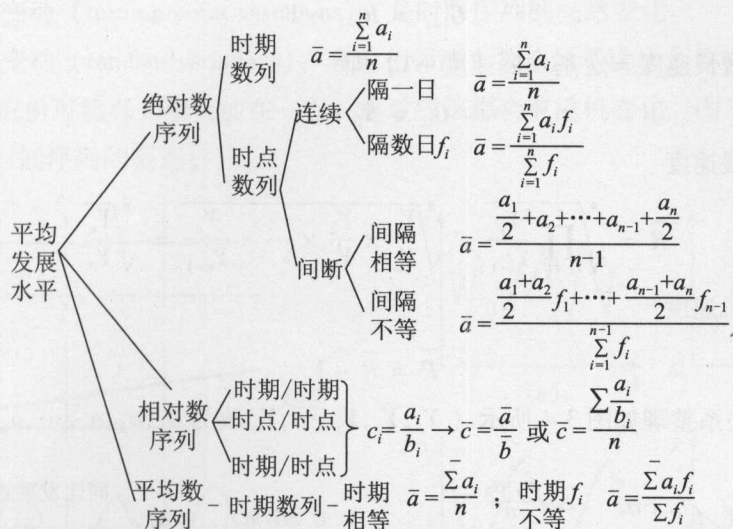


图 3-3 水平分析

1. 发展速度

(1) 同比发展速度: 和去年同期对比, $R_{i-12}^i = \frac{Y_i}{Y_{i-12}}$ 。

(2) 环比发展速度: 和上一期对比, $R_{i-1}^i = \frac{Y_i}{Y_{i-1}}$ 。环比发展速度是第 4 章指数, 个体指数的环比指数。

(3) 定基发展速度: 和基期对比, $R_0^i = \frac{Y_i}{Y_0}$ 。定基发展速度是第 4 章指数, 个体指数的定基指数。

环比发展速度和定基发展速度满足以下关系。

(1) 环比 \times 环比 $\times \dots \times$ 环比 = 定基, 即

$$\prod_{i=1}^n R_{i-1}^i = R_0^n, \prod_{i=1}^n \frac{Y_i}{Y_{i-1}} = \frac{Y_n}{Y_0}$$

(2) 定基 \div 定基 = 环比, 即

$$\frac{R_0^i}{R_0^{i-1}} = R_{i-1}^i, \frac{Y_i}{Y_0} \div \frac{Y_{i-1}}{Y_0} = \frac{Y_i}{Y_{i-1}}$$

2. 增长速度

(1) 同比增长速度 = 同比发展速度 - 1, 即

$$G_{i-12}^i = R_{i-12}^i - 1$$

(2) 环比增长速度 = 环比发展速度 - 1, 即

$$G_{i-1}^i = R_{i-1}^i - 1$$

(3) 定基增长速度 = 定基发展速度 - 1, 即

$$G_0^i = R_0^i - 1$$

3. 平均发展速度

$$\bar{R} = \sqrt[n]{\prod_{i=1}^n \frac{Y_i}{Y_{i-1}}} = \sqrt[n]{\frac{Y_1}{Y_0} \times \frac{Y_2}{Y_1} \times \cdots \times \frac{Y_n}{Y_{n-1}}} = \sqrt[n]{\frac{Y_n}{Y_0}}$$

4. 平均增长速度

$$\bar{P} = \bar{R} - 1$$

以上式子关系整理如图 3-4 所示 ($Y_1, Y_2, Y_3, \cdots, Y_n$ 改为 $a_1, a_2, a_3, \cdots, a_n$)。

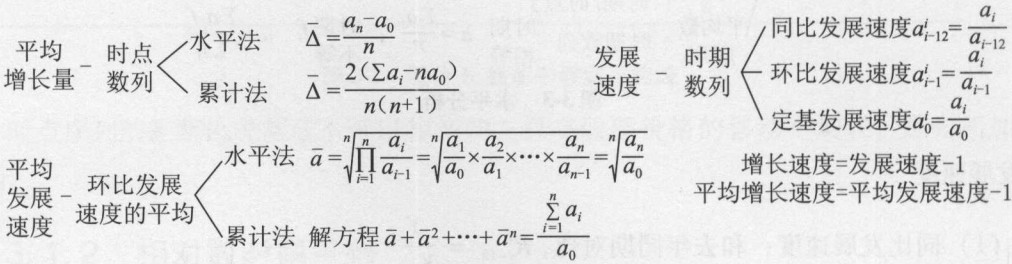


图 3-4 速度分析

例题 3.1 水平分析与速度分析。(解答见网络资源)

3.3 时间序列构成因素

从本节开始, 以下时间序列分析, 适用在绝对数的“时期序列”。

在时间序列分析, 历史数据可以分解成 6 个成份。

(1) 水平 (levels): 表示时间序列有固定的数值。

(2) 趋势 (trends): 表示时间序列有斜率的上升或下降。

(3) 季节变动 (seasonal variations): 周期性的变化, 通常是每年一个周期, 因为气候、季节、假期等因素而有周期性。例如下列产品的需求会有季节变化: 冷气机、冰淇淋、毛衣、教科书、卡片等。

(4) 循环变动 (cyclical variations): 是一种长时期的振动或摆动, 其间隔均在一年以上, 通常是经济活动影响企业扩张或萎缩的循环, 例如: 经济景气、萧条或低迷等循环变动。

(5) 转捩变动 (turning point variations): 是时间序列的突然变化。

(6) 随机变动 (random variations): 表示相邻两个数值的干扰、余数或误差。

以上各种成份可能有几项会加在一起, 不过应该都会有随机变化。图 3-5 是 6 个成份, 图 3-6 是一些时间序列的型态。

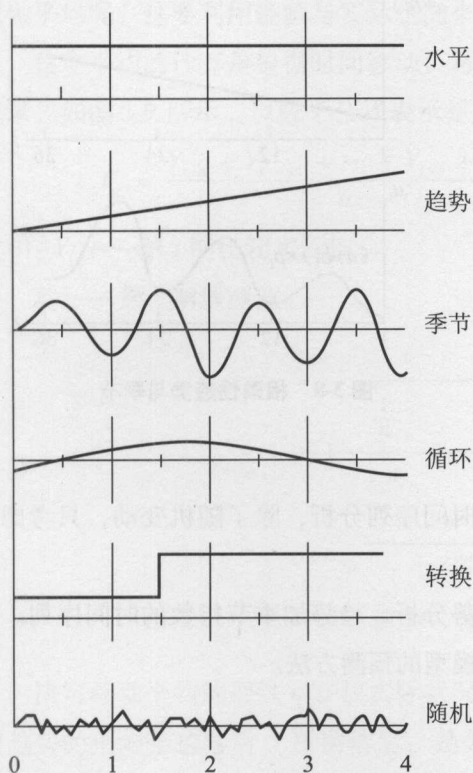


图 3-5 时间序列的成份

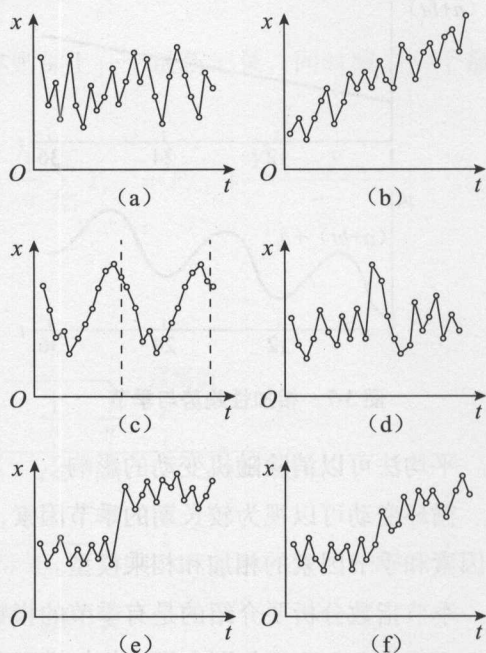


图 3-6 时间序列型态

(a) 常数或水平; (b) 线性趋势; (c) 循环变动; (d) 遽增; (e) 阶梯; (f) 倾斜
注: 以上均加上随机变动。

时间序列的指针符号为 T 、 S 、 C 、 I , 分别表示趋势、季节、循环和随机, 时间序列分析模型有下列两种模型。

(1) 加法模型 (additive trend seasonal): 季节的振幅并不跟着销售额增加而变大, 如图 3-7 所示, 其关系式是

$$Y = T + S + C + I$$

销售需求 = 趋势因素 + 季节因素 + 循环变动 + 随机变动

(2) 乘法模型 (multiplicative trend seasonal): 根据销售额增加, 季节的振幅也摆动越大, 如图 3-8 所示, 其关系式是

$$Y = T \times S \times C \times I$$

销售需求 = 趋势因素 \times 季节因素 \times 循环变动 \times 随机变动

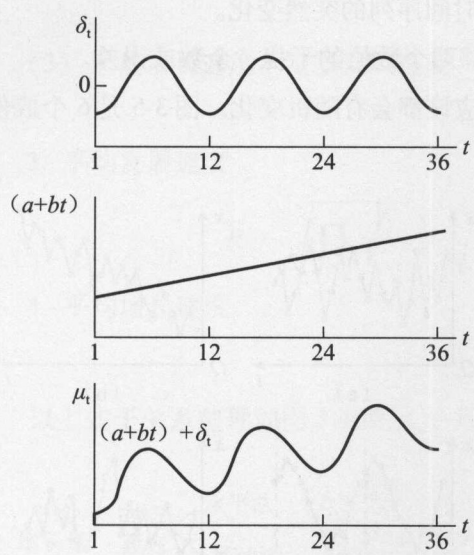


图 3-7 相加性趋势与季节

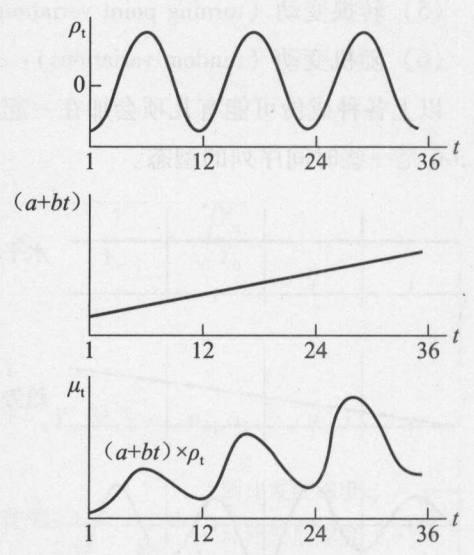


图 3-8 相乘性趋势与季节

平均法可以消除随机变动的影响。

循环变动可以视为较长期的季节因素，所以时间序列分析，除了随机变动，只考虑趋势因素和季节因素的相加和相乘模型。

季节指数分析要介绍的是有季节的指数和趋势分析，趋势加季节指数的时间序列。本章 3.6 节和 3.7 节将分别介绍，相加模型和相乘模型的预测方法。

一般时间序列分析多采用相乘模型。

3.4 平稳型序列预测

平稳型序列是只有水平和随机变动，预测主要是消除随机。预测方法有：简单移动平均法、加权移动平均、简单指数平滑法。

3.4.1 简单移动平均法

简单移动平均法 (simple moving average) 是将最近 k 期的实际值加以平均，来预测下一期的实际值。股票价格预测的技术分析之一，就是利用不同期数的移动平均数。

股价的移动平均线，简称均线，就是将 k 天的收盘价总和再除以 k ，得到第 k 天的算术平均线数值。5 日均线是过去 5 天收盘价的平均值，又称为周线 (5 天交易)；20 日均线又称月线，是多头股票的防守点。短期均线在长期均线之上的话，就属一种多头走势，

反之则是空头走势之一。但是，股价序列是时点序列，移动平均的均线，适合买卖股票的技术分析，不适合下列的预测方法。

移动平均的目的是要包括足够的期数来消除随机变化。但是期数也不要太多，如此可使过去不重要的信息不必计算，同时也不需要保留太多的记录。到底要选择多少期（ k ）来做平均呢？这要利用经验与实际状况来决定。

移动平均的计算是根据时间移动，每次计算则加上一个最新记录，同时减少一个最旧记录，如图 3-9 所示。以数学公式表示是

$$F_t = \frac{Y_{t-n} + Y_{t-n+1} + \cdots + Y_{t-2} + Y_{t-1}}{n} = \frac{1}{n} \sum_{i=1}^n Y_{t-i} = F_{t-1} + \frac{Y_{t-1} - Y_{t-n-1}}{n}$$

式中： Y_i ——第 i 期已知实际值；

F_t ——第 t 期预测值。

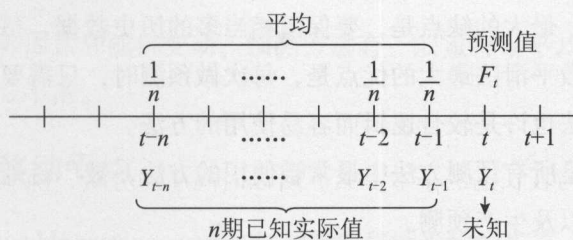


图 3-9 简单移动平均法计算

决定移动平均的期数 n 要视实际状况而定。如果期数 n 越大，则预测值越平稳，但是对趋势的预测会越落后（所谓落后，是如果趋势是上升，则预测值偏低；如果趋势是下降，则预测值偏高）。如果期数 n 越小，则对趋势之预测的落后会越小，但是对随机变化的结果越不平稳。

如果要利用简单移动平均的话，那么期数 n 可以选择小一点的。如果时间序列的随机变化大，则期数 n 可以选大一点的。较适当的 n 的决定，可以利用本章 3.9 节预测误差，找出使预测误差最小的 n 。

总之，简单移动平均可以消除随机成份（使误差减小），而且可以表示趋势（虽然会有一些落后），但是不能解决季节变化。

例题 3.2 简单移动平均法。（解答见网络资源）

3.4.2 加权移动平均

如果时间序列有趋势成份，则简单移动平均取期数 n 小一点，会使趋势不致落后太多，但是却失去了早期的实际值。

为了解决这个问题,加权移动平均在 n 期实际值中分别给予不同的权数,再加以平均。通常越近期实际值的权数越大,越早期实际值的权数越小。简单移动平均是每期的权数均相同。加权移动平均 (weighted moving average) 的公式表示如下

$$F_t = (W_{t-n}Y_{t-n} + W_{t-n+1}Y_{t-n+1} + \cdots + W_{t-2}Y_{t-2} + W_{t-1}Y_{t-1}) / (\sum_{i=1}^n W_{t-i})$$

$$= \sum_{i=1}^n W_{t-i}Y_{t-i} / \sum_{i=1}^n W_{t-i}$$

式中: Y_t ——第 t 期的实际值;

F_t ——第 t 期的预测值;

W_t ——第 t 期的权数, $W_{t-n} \leq W_{t-n+1} \leq \cdots \leq W_{t-2} \leq W_{t-1}$ 。

3.4.3 简单指数平滑法

在移动平均法中,最大的缺点是,要保留相当多的历史数据,每次新数据加入,才能丢掉最老的数据。指数平滑法最大的优点是,每次做预测时,只需要上一期的真实值与预测值。所以指数平滑法也许是较合逻辑而容易使用的方法。

指数平滑法可能是所有预测方法中最常被使用的方法,被广泛地应用于批发,零售与服务业的存货管理,以及生产预测。

简单指数平滑 (simple exponential smoothing) 只需要 3 个数据做预测: 上期实际值, 上期预测值, 以及平滑常数 α (smoothing constant)。平滑常数决定平滑的程度, 以及反映实际值与预测值之差。平滑常数越小, 则预测值越平滑; 平滑常数越大, 则越反映实际值与预测值之差。平滑常数的决定, 在于产品的需求特性及管理者的感觉判断。例如, 生产标准产品而有相当稳定的需求, 则平滑常数可以很小。如果产品需求具有相当成长或高度变动, 则可能需要较高的平滑常数。一般 α 介于 0.01 与 0.3 之间, 而 $\alpha = 0.1$ 是合理的平滑常数。

简单指数平滑的模型是

$$F_t = F_{t-1} + \alpha(Y_{t-1} - F_{t-1}) = \alpha Y_{t-1} + (1 - \alpha)F_{t-1}$$

式中: F_t ——第 t 期的预测值;

F_{t-1} ——第 $t-1$ 期的预测值;

Y_{t-1} ——第 $t-1$ 期的实际值;

α ——平滑常数, $0 \leq \alpha \leq 1$ 。

$$F_{t+1} = \alpha Y_t + \alpha(1 - \alpha)Y_{t-1} + \alpha(1 - \alpha)^2 Y_{t-2} + \cdots + \alpha(1 - \alpha)^{t-1} Y_1 + (1 - \alpha)^t F_0$$

所以, 简单指数平滑法是一种加权移动平均, 越接近 t 期的权数越大。

起始预测值是先找 n 个值作平均: $F_0 = \sum Y_i / n$ 。

简单指数平滑的步骤是：

第1步：决定平滑常数 α 。

第2步：假设有 T 个实际值，从这当中找最初的 n 个数做平均，令 $F_{n+1} = F$ 起始值。

第3步： $F_{n+2} = \alpha Y_{n+1} + (1 - \alpha)F_{n+1}$ 。

第4步：依此求到 $F_{T+1} = \alpha Y_T + (1 - \alpha)F_T$ 。

第5步： T 期以后的预测值均等于 F_{T+1} ，即 $F_{T+k} = F_{T+1}$ 。

例题 3.3 简单指数平滑。（解答见网络资源）

3.5 趋势型序列预测

趋势型序列有趋势因素和随机变动，预测方法有：双重移动平均、一元线性回归、双重指数平滑、趋势指数平滑。

3.5.1 双重移动平均

双重移动平均（double moving average）是简单移动平均的扩充（如图 3-10 所示），主要目的是用于有趋势的时间序列。双重移动平均有两个移动平均值 M_t 与 $M_t^{(2)}$ 。假设目前是在第 t 期，已知实际值 Y_t ，则

$$M_t = (Y_{t-n+1} + \cdots + Y_{t-1} + Y_t)/n = M_{t-1} + (Y_t - Y_{t-n})/n, t \geq k$$

$$M_t^{(2)} = (M_{t-n+1} + \cdots + M_{t-1} + M_t)/n = M_{t-1}^{(2)} + (M_t - M_{t-n})/n, t \geq 2k - 1$$

$$a_t = 2M_t - M_t^{(2)}$$

$$b_t = 2(M_t - M_t^{(2)})/(n - 1)$$

所以第 $t+k$ 期的预测值 $F_{t+k} (k \geq 1)$ 为

$$F_{t+k} = a_t + b_t \times k$$

式中： Y_t ——已知实际值；

F_{t+k} ——第 $t+k$ 期之预测值；

M_t ——简单移动平均值；

$M_t^{(2)}$ ——2 次移动平均值；

a_t ——趋势直线之常数；

b_t ——趋势直线之斜率。

所以简单移动平均与加权移动平均只能预测下一期之值（如果要预测第 $t+k$ 期，则

$F_{t+k} = F_{t+1}, k \geq 1$); 而双重移动平均可以预测以后数期的值。

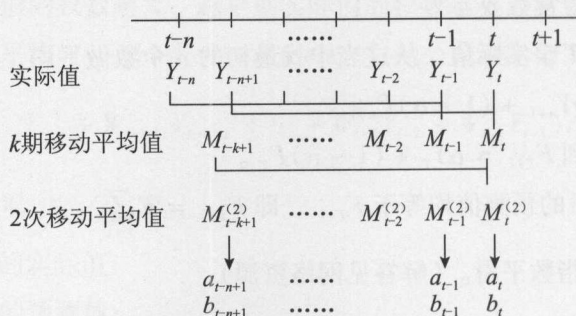


图 3-10 双重移动平均之计算

例题 3.4 双重移动平均。(解答见网络资源)

3.5.2 一元线性回归

一元线性回归 (simple linear regression) 又称方程拟合法, 一元的意思是指有一个自变量, 线性的意思是因变量和自变量, 存在线性关系。时间序列法的线性回归和因果关系法的回归分析 (请见第 13 章) 是相同的, 只是这里的线性回归不用其他外在独立变量, 而是用时间当作独立变量。换言之, 我们是假设, 要预测的销售量与时间有线性关系, 即

$$\hat{Y}_t = a + bt \quad t = 1, 2, \dots, n$$

$$b = (n \sum_{t=1}^n tY_t - \sum_{t=1}^n t \sum_{t=1}^n Y_t) / [n \sum_{t=1}^n t^2 - (\sum_{t=1}^n t)^2]$$

$$a = (\sum_{t=1}^n Y_t - b \sum_{t=1}^n t) / n$$

$$F_t = a + bt$$

$$S^2 = (\sum_{t=1}^n Y_t^2 - a \sum_{t=1}^n Y_t - b \sum_{t=1}^n tY_t) / (n - 2)$$

$$r^2 = (n \sum_{t=1}^n tY_t - \sum_{t=1}^n t \sum_{t=1}^n Y_t)^2 / \{ [n \sum_{t=1}^n t^2 - (\sum_{t=1}^n t)^2] [n \sum_{t=1}^n Y_t^2 - (\sum_{t=1}^n Y_t)^2] \}$$

式中: \hat{Y}_t ——第 t 期的回归预测值;

a ——回归直线的估计常数;

b ——回归直线的估计斜率;

F_t ——第 t 期的预测值;

S ——回归预测值的标准差;

r ——回归相关系数;

Y_t ——第 t 期的实际值。

若 $0.9 \leq r \leq 1$ ，则实际值和时间有很高的相关；若 $0.7 \leq r \leq 0.89$ ，则是高相关；若 $0.4 \leq r \leq 0.69$ ，则有中度相关；若 $0.2 \leq r \leq 0.39$ ，则是低相关；若 $0 \leq r \leq 0.19$ ，则有很低的相关。相关性越高，则预测值越准确。

利用标准差，可以计算预测范围，即 Y_t 在 $F_t \pm S$ 之内有 68% 之概率； Y_t 在 $F_t \pm 2S$ 之内有 95% 之概率。

例题 3.5 一元线性回归。（解答见网络资源）

3.5.3 双重指数平滑

双重指数平滑（double exponential smoothing），以直线方程式 $f(t) = a + bt$ 来做预测。现在我们要利用过去预测值与实际值来估计 a 和 b ，即

$$S_t = \alpha Y_t + (1 - \alpha) S_{t-1}$$

$$S_t^{(2)} = \alpha S_t + (1 - \alpha) S_{t-1}^{(2)}$$

$$a_t = 2S_t - S_t^{(2)}$$

$$b_t = \frac{\alpha}{1 - \alpha} (S_t - S_t^{(2)})$$

$$F_{t+k} = a_t + b_t \times k$$

式中： S_t ——第 t 期简单指数平滑值；

α ——平滑常数， $0 \leq \alpha \leq 1$ ；

Y_t ——第 t 期的实际值；

$S_t^{(2)}$ ——第 t 期二次指数平滑值；

a_t ——第 t 期趋势直线常数；

b_t ——第 t 期趋势直线斜率；

F_{t+k} ——第 $t+k$ 期的预测值。

双重指数平滑的步骤是：

第 1 步：决定平滑系数 α 。

第 2 步：假设有 T 个实际值，从这当中找最初 n 个值做直线回归。

$Y_t = a + bt$ ，得到估计值 a_0 和 b_0

$$S_0 = a_0 - \left(\frac{1 - \alpha}{\alpha} \right) b_0$$

$$S_0^{(2)} = a_0 - 2 \left(\frac{1 - \alpha}{\alpha} \right) b_0$$

第3步：计算

$$S_{n+1} = \alpha Y_{n+1} + (1 - \alpha) S_0$$

$$S_{n+1}^{(2)} = \alpha S_{n+1} + (1 - \alpha) S_0^{(2)}$$

$$a_{n+1} = 2S_{n+1} - S_{n+1}^{(2)}$$

$$b_{n+1} = \frac{\alpha}{1 - \alpha} (S_{n+1} - S_{n+1}^{(2)})$$

$$F_{n+2} = a_{n+1} + b_{n+1}$$

第4步：依此类推，求 $S_T, S_T^{(2)}, a_T, b_T$ ，则 $F_{T+1} = a_T + b_T$ 。

第5步： T 以后的预测值为 $F_{T+k} = a_T + b_T \times k$ 。

3.5.4 趋势指数平滑

趋势指数平滑 (trend exponential smoothing)，是将趋势线加到时间序列数据中一起考虑。而趋势线的估计方法则是考虑每个过去上升或下滑的增量或减量，并以最近的时间序列数据作计算最近一期趋势的依据。

以趋势指数平滑来做预测，必须利用过去预测值与实际值来估计趋势线和平滑曲线。首先预测估计趋势值 (estimated trend)

$$LT_t = \alpha(Y_t - Y_{t-1}) + (1 - \alpha)(F_t - F_{t-1})$$

$$ET_{t+1} = \beta(LT_t) + (1 - \beta)(ET_t)$$

式中： LT_t ——第 t 期之最近趋势值；

α ——平滑常数， $0 \leq \alpha \leq 1$ ；

Y_t ——第 t 期的实际值；

ET_t ——第 t 期之估计趋势值；

β ——趋势平滑常数， $0 \leq \beta \leq 1$ ；

F_t ——第 t 期的预测值。

从这个计算公式来看，估计趋势值 (estimated trend) 是在计算过去最近一期与过去最近两期间的上升或下滑的增量或减量，用其与本期估计趋势值来估计下一期的趋势，其概念与指数平滑的概念类似，只是将原用于时间序列数据的方法再应用于趋势的计算。

必须注意的是以趋势指数平滑法来预测，使用者得先给定两个值介于 0 到 1 之间的指数，一为平滑常数 α ，另一为趋势平滑常数 β ，顾名思义， α 之用法与指数平滑的用法类似，而 β 则是用于预测估计趋势值。

计算估计趋势值 (ET) 之后，就可以用趋势指数平滑法来计算未来的预测值 (F_t)

$$F_{t+1} = \alpha Y_t + (1 - \alpha) F_t + ET_{t+1}$$

式中： α ——平滑常数， $0 \leq \alpha \leq 1$ ；

Y_t ——第 t 期的实际值；

F_t ——第 t 期的预测值；

ET_{t+1} ——第 $t+1$ 期之估计趋势值。

未来一期预测值的算法与指数平滑一样，但是最后加入估计趋势值，除可以将未来趋势考虑进去外，也可解决指数平滑只能预测出过去一中间值的现象。

但是这个方法的运用得当与否，取决于起始值的设定。起始值的决定是先找最初的实际值（ Y_0 ）、预测值（ F_0 ）与估计趋势值（ ET_1 ），通常会设定为

$$Y_0 = Y_1 \text{ 或 } \bar{Y}_t$$

$$F_0 = Y_0$$

$$ET_1 = 1 \text{ 或 } \sum (Y_t - Y_{t-1}) / (n - 1) \text{ (如果总共有 } n \text{ 期数据)}$$

因此第一期的预测值（ F_1 ）： $F_1 = Y_0 + ET_1$ 。

由于趋势指数平滑法来预测加入考虑未来趋势，如果总共有 n 期数据，那么当预测未来第 k 期之后的预测值（ F_{n+k} ）就可以用以下方式进行计算

$$F_{n+k} = \alpha Y_n + (1 - \alpha) F_n + k(ET_{n+1})$$

趋势指数平滑法的步骤是：

第1步：决定最初值（ Y_0 ），估计趋势值（ ET_1 ），平滑常数 α ，趋势平滑常数 β 。

第2步：计算第一期的预测值 $F_1 = Y_0 + ET_1$ 。

第3步：假设有 n 个实际值，计算第二期估计趋势值（ ET_2 ）

$$LT_t = \alpha(Y_t - Y_{t-1}) + (1 - \alpha)(F_t - F_{t-1})$$

$$ET_{t+1} = \beta(LT_t) + (1 - \beta)(ET_t)$$

用趋势指数平滑法来计算未来的预测值（ F_t ）

$$F_{t+1} = \alpha Y_t + (1 - \alpha) F_t + ET_{t+1}$$

第4步：依此类推，计算未来 $n-2$ 期之估计趋势值（ ET_t ）与预测值（ F_t ）。

第5步： n 期以后的第 k 期预测值为 $F_{n+k} = \alpha Y_n + (1 - \alpha) F_n + k(ET_{n+1})$ 。

趋势指数平滑法计算步骤如表 3-1 所示。

表 3-1 趋势指数平滑法计算步骤

t	Y_t	ET_t	F_t	LT_t
0	Y_0		F_0	
1	Y_1	ET_1	$F_1 = Y_0 + ET_1$	$\alpha(Y_1 - Y_0) + (1 - \alpha)(F_1 - F_0)$

续表

2	Y_2	$\beta(LT_1) + (1 - \beta)(ET_1)$	$\alpha Y_1 + (1 - \alpha)F_1 + ET_2$	$\alpha(Y_2 - Y_1) + (1 - \alpha)(F_2 - F_1)$
\vdots	\vdots	\vdots	\vdots	\vdots
n	Y_n	$\beta(LT_{n-1}) + (1 - \beta)(ET_{n-1})$	$\alpha Y_{n-1} + (1 - \alpha)F_{n-1} + ET_n$	$\alpha(Y_n - Y_{n-1}) + (1 - \alpha)(F_n - F_{n-1})$
$n+1$		$\beta(LT_n) + (1 - \beta)(ET_n)$	$\alpha Y_n + (1 - \alpha)F_n + ET_{n+1}$	

例题 3.6 趋势指数平滑。(解答见网络资源)



3.6 季节指数分析

季节性指数计算有两种方法：中央移动平均法（centered moving averages）和趋势方程拟合法或一元线性回归法（simple linear regression），如图 3-11 所示。

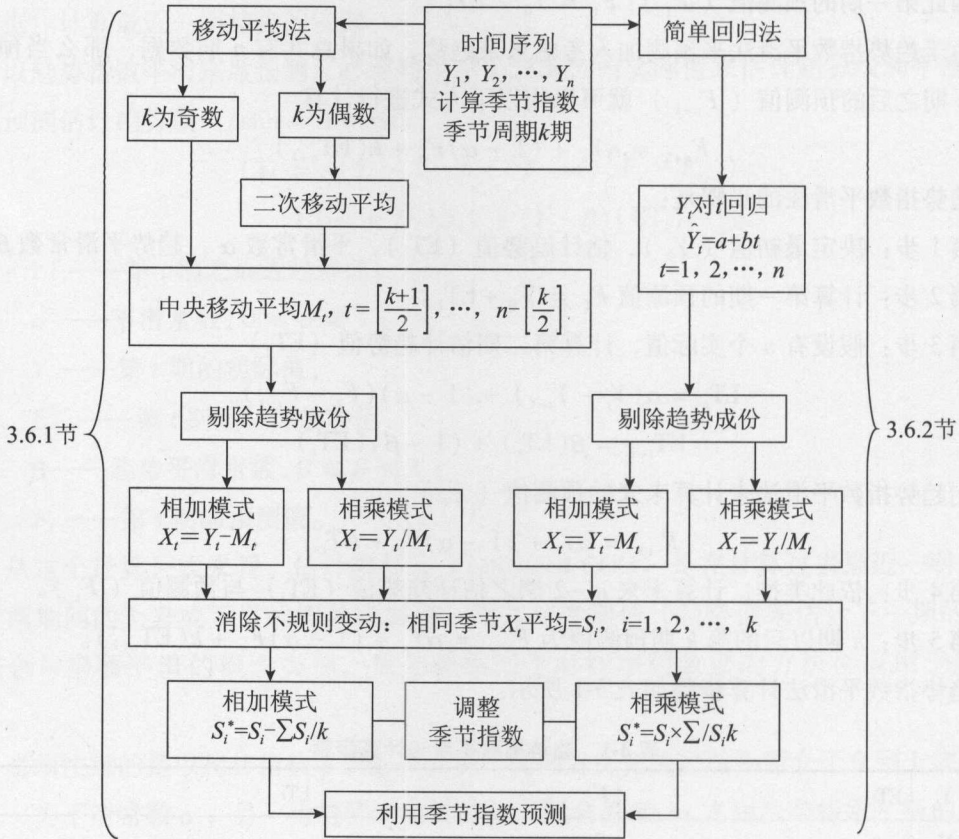


图 3-11 计算季节性指数

3.6.1 中央移动平均法计算季节性指数

利用中央移动平均法 (centered moving averages) 分解时间数列, 是找出时间数列中的趋势及季节系数 (或循环系数)。

假设时间数列是 $\{Y_1, Y_2, Y_3, \dots, Y_n\}$, 其计算步骤如下。

第1步: 利用观察法找出季节的周期长度 (或循环的周期长度) k 。

第2步: 以 k 期计算移动平均。

(1) 若 k 是奇数且 $t > (k-1)/2, t = (k+1)/2, \dots, n - (k-1)/2$, 则

$$M_t = [Y_{t-(k-1)/2} + Y_{t-(k-3)/2} + \dots + Y_{t-1} + Y_t + Y_{t+1} + \dots + Y_{t+(k-3)/2} + Y_{t+(k-1)/2}] / k$$

如果 k 为偶数, 则相邻两个移动平均再取平均。

二次移动平均: 第一次以 k 期计算移动平均, 第二次以 2 期计算移动平均。

(2) 若 k 是偶数且 $t > k/2, t = k/2 + 1, \dots, n - k/2$, 则

$$M_t = (Y_{t-k/2} + 2Y_{t-k/2+1} + \dots + 2Y_{t-1} + 2Y_t + 2Y_{t+1} + \dots + 2Y_{t+k/2-1} + Y_{t+k/2}) / 2k$$

例如: $k=5$, 则 $M_t = (Y_{t-2} + Y_{t-1} + Y_t + Y_{t+1} + Y_{t+2}) / 5, t > 2, t = 3, \dots, n-2$

$$M_3 = (Y_1 + Y_2 + Y_3 + Y_4 + Y_5) / 5$$

例如: $k=4$, 则 $M_t = (Y_{t-2} + 2Y_{t-1} + 2Y_t + 2Y_{t+1} + Y_{t+2}) / 8, t > 2, t = 3, \dots, n-2$

$$M_{2.5} = (Y_1 + Y_2 + Y_3 + Y_4) / 4, M_{3.5} = (Y_2 + Y_3 + Y_4 + Y_5) / 4 \Rightarrow$$

$$M_3 = (M_{2.5} + M_{3.5}) / 2 = (Y_1 + 2Y_2 + 2Y_3 + 2Y_4 + Y_5) / 8$$

请注意, 趋势成份与预测的移动平均公式不同, 例如以 3 期做移动平均 $k=3$:

做预测的简单移动平均公式

$$F_5 = (Y_2 + Y_3 + Y_4) / 3$$

双重移动平均公式

$$M_5 = (Y_3 + Y_4 + Y_5) / 3$$

计算趋势与季节成份的中央移动平均公式

$$M_5 = (Y_4 + Y_5 + Y_6) / 3$$

第3步: 移动平均 M_t 为趋势系数, $t = \lceil (k+1)/2 \rceil, \dots, n - \lfloor k/2 \rfloor$ 。

“ x ”是大于等于 x 的最小整数, 即 x 若有大于 0 的小数, 则去掉小数, 整数加 1。

$\lfloor x \rfloor$ 是小于等于 x 的最大整数, 即 x 若有大于 0 的小数, 则去掉小数, 保留整数。

例如: 若 $n=15, k=4$, 则 $M_t, t = 3, \dots, 13$; 若 $n=18, k=9$, 则 $M_t, t = 5, \dots, 14$ 。

第4步: 季节系数, 如果是相加性模式, 则 $X_t = Y_t - M_t, t = \lceil (k+1)/2 \rceil, \dots, n - \lfloor k/2 \rfloor$ 。

如果是相乘性模式, 则 $X_t = Y_t / M_t, t = \lceil (k+1)/2 \rceil, \dots, n - \lfloor k/2 \rfloor$ 。

第5步：季节指数，同一季的 X_t 取平均数为季节指数，依照其所属之季节，分别算出平均值 S_i 。

$$S_i = (X_i + X_{i+k} + X_{i+2k} + \cdots + X_{i+mk}) / (m + 1)$$

第6步：调整季节指数，将此平均值 S_i 做正规化。

相加性模式，将每个 S_i 减去 S_i 的平均值： $S_i^* = S_i - \sum_{i=1}^k S_i / k$ ，因此 $\sum_{i=1}^k S_i^* = 0$ 。

相乘性模式，将每个 S_i 乘上 k 再除以 S_i 的总和： $S_i^* = S_i \times k / \sum_{i=1}^k S_i$ ，因此 $\sum_{i=1}^k S_i^* = k$ 。

第7步：利用季节指数，进行预测，请见3.7节及3.8节。

例题 3.7 以中央移动平均法计算季节指数。（解答见网络资源）

3.6.2 趋势方程拟合法

趋势方程拟合法就是直线回归法，计算季节性指数时，是将直线回归法所做出的直线当成是时间序列比较的基准，也就是如同中央移动平均法所做出的长期趋势值。

计算直线回归线时，将原始时间序列数据当成因变量，时间则为自变量，计算出直线回归线值之后就可以开始计算季节性指数，并以此指数做未来时间序列预测，其步骤与之前以中央移动平均法计算季节性指数的步骤雷同，如下：

第1步：以时间序列数据 Y_t 为因变量，时间 t 为自变量，计算直线回归线。

第2步：计算季节系数，如果是相加性模式，则 $X_t = Y_t - \hat{Y}_t$ ， $t = 1, 2, \cdots, n$ 。

如果是相乘性模式，则 $X_t = Y_t / \hat{Y}_t$ ， $t = 1, 2, \cdots, n$ 。

第3步：季节指数，同一季的 X_t 取平均数为季节指数，依照其所属之季节，分别算出平均值 $S_i = \sum_{j=1}^a X_{i+jk} / a$ ， $i = 1, \cdots, k$ 。

第4步：调整季节指数，将此平均值 S_i 做正规化（normalization）。

相加性模式，将每个 S_i 减去 S_i 的平均值： $S_i^* = S_i - \sum_{i=1}^k S_i / k$ ，因此 $\sum_{i=1}^k S_i^* = 0$ 。

相乘性模式，将每个 S_i 乘上 k 再除以 S_i 的总和： $S_i^* = S_i \times k / \sum_{i=1}^k S_i$ ，因此 $\sum_{i=1}^k S_i^* = k$ 。

例题 3.8 趋势方程拟合法。（解答见网络资源）

得到季节指数以后，就可以进行时间序列的预测。分别有趋势加季节的相加模式和趋势乘季节的相乘模式，如图3-12所示，在下两节说明。

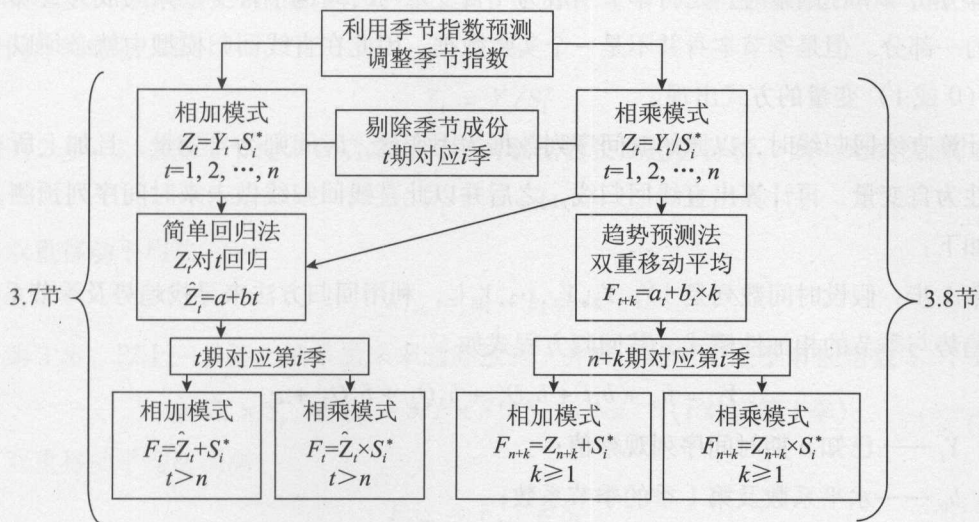


图 3-12 利用季节指数预测

3.7 时间序列预测方法：趋势加季节

3.7.1 相加性季节指数

第1步：将时间序列值减去相对应之季节性指数 S_i^* ，计算出每期去除季节性因素后之时间序列数据，即

$$Z_t = Y_t - S_i^*$$

第2步：以一元线性回归做长期趋势分析，其中以去除季节性因素后之时间序列数据 Z_t 为因变量，以 t 为自变量。回归直线为

$$\hat{Z}_t = a + bt$$

第3步：上一步骤之结果做未来趋势预测，并加上或乘上相对应之季节性指数 S_i^* ，即

$$F_t = \hat{Z}_t + S_i^* = (a + bt) + S_i^* \text{ 或 } F_t = \hat{Z}_t \times S_i^* = (a + bt) \times S_i^*, t > n \text{ (} t \text{ 对应第 } i \text{ 季)}$$

例题 3.9 相加性季节指数。(解答见网络资源)

3.7.2 利用多元回归方法

相加性季节性趋势分析模型，其关系式是：销售需求 = 趋势 + 季节因素。以直线回归

法计算分析季节性因素时,是将季节当成为一自变量与时间这个自变量共同成为直线回归模型的一部分。但是季节本身并不是一个实数数据,因此在直线回归模型中就必须以一个 0-1 (0 或 1) 变量的方式出现。

计算直线回归线时,以原始时间序列数据为因变量,时间则为自变量,且加上所有的季节性为自变量,再计算出直线回归线,之后并以此直线回归线做未来时间序列预测,其步骤如下:

第 1 步:假设时间数列是 $\{Y_1, Y_2, Y_3, \dots, Y_n\}$, 利用回归方法来寻找趋势及季节系数。趋势与季节的相加性模式,其回归方程式如下

$$Y_t = b_0 + b_1 t + b_2 Q_1 + b_3 Q_2 + b_4 Q_3 + \varepsilon_t$$

式中: Y_t ——已知 n 期时间序列观察值;

b_0 ——水平系数及第 1 季的季节系数;

b_1 ——趋势系数;

b_2, b_3, b_4 ——第 1, 2, 3 季的季节系数;

ε_t ——残差项,实际值与预测值之差;

Q_i ——第 i 季的季节变量, $i=1, 2, 3$ 。

若 Y_t 是在第 i 季, $i=1, 2, 3$, 则 $Q_i=1$; 若 Y_t 不是在第 i 季, $i=1, 2, 3$, 则 $Q_i=0$; 若数据是在第 4 季, 则 $Q_1 = Q_2 = Q_3 = 0$ 。

Q_i 是虚拟变量 (dummy variable), Q_i 为分类尺度, 当有 k 类 (例如 4 季), 则只要 $k-1$ 个虚拟变量。

Y_t 对 t, Q_1, Q_2, Q_3 多元回归, 得到回归方程

$$\hat{Y}_t = b_0 + b_1 t + b_2 Q_1 + b_3 Q_2 + b_4 Q_3$$

第 2 步: 对于 n 期以后的未知值 $Y_t, t > n$, 其预测值 $F_t, t > n$

$$F_t = \hat{Y}_t = b_0 + b_1 t + b_2 Q_1 + b_3 Q_2 + b_4 Q_3$$

利用回归方法分解时间数列, 是找出时间数列中的趋势及季节系数。

例题 3.10 多元回归方法。(解答见网络资源)

3.8 时间序列预测方法: 趋势乘季节

3.8.1 季节指数预测

季节指数预测的计算步骤:

第1步：将时间序列值除以相对应之季节性指数 S_i^* ，计算出每期去除季节性因素后之时间序列数据，即

$$Z_t = Y_t / S_i^*$$

第2步：以去除季节性因素后之时间序列数据做长期趋势分析，如一元线性回归

$$\hat{Z}_t = a + bt$$

双重移动平均法

$$F_{t+k} = a_t + b_t \times k$$

第3步：以上一步骤之结果做未来趋势预测，并乘上相对应之季节性指数 S_i^* ，即

$$F_t = \hat{Z}_t \times S_i^* = (a + bt) \times S_i^*, \quad t > n \quad (t \text{ 对应第 } i \text{ 季})$$

双重移动平均法预测

$$F_{n+k} = F_{n+k} \times S_i^*$$

例题 3.11 季节指数预测。(解答见网络资源)

3.8.2 季节指数平滑

季节指数平滑法 (seasonal exponential smoothing) 要介绍的是有趋势和季节的指数平滑，也就是 Winters 法 (Winters' model)。

Winters 法是假设趋势与季节相乘性效果，需要 3 个平滑常数 (α, β, γ)。建议平滑常数 $\alpha = 0.2, \beta = 0.05, \gamma = 0.1$ 。假设 L 是季节循环周期，例如以月计算，则 $L = 12$ ；以季计算，则 $L = 4$ 。计算公式为

$$a_t = \alpha \left(\frac{Y_t}{c_{t-L}} \right) + (1 - \alpha)(a_{t-1} + b_{t-1})$$

$$b_t = \beta(a_t + a_{t-1}) + (1 - \beta)b_{t-1}$$

$$c_t = \gamma \left(\frac{Y_t}{a_t} \right) + (1 - \gamma)c_{t-L}$$

$$F_{t+k} = (a_t + b_t k) c_{t+k-L}$$

式中： a_t ——第 t 期趋势直线估计值；

Y_t ——第 t 期的实际值；

c_t ——第 t 期季节系数；

b_t ——第 t 期趋势直线斜率；

α, β, γ ——平滑常数；

F_{t+k} ——第 $t+k$ 期的预测值。

3.9 预测误差

本章所讨论的时间序列预测方法适合各种状况，但是这些方法并没有真正的优劣势存在，也没有任何的方法特别准确，使用时由于有中文统计或其他软件辅助，都可以很快地算出结果。因此建议针对每个问题用各种方法测试，以下所讨论的误差计算方法，比较看哪个方法较为准确，再选用该方法做预测用。

3.9.1 误差来源

误差的来源有许多种，通常的来源是由过去数据统计出来的。例如，利用回归分析，我们可以得到统计误差，也就是回归直线与实际值之距离平方的平均，从统计误差可以得到预测上下限的范围。但是当我们实际预测未来时，有可能又超出这个预测范围，因为预测范围只是根据过去数据计算的，并不能完全说明未来的结果。所以，实际误差可能超出预测模型所估计的统计误差。

误差可以分成偏误 (bias) 及随机 (random) 误差。偏误是因为错误运用模型所致。偏误的来源有：没有包括适当的变量，变量间使用错误的关系，利用错误的趋势线，以及没有利用趋势与季节因素等。偏误类似第 1 章的非抽样误差，非抽样误差针对数据，偏误针对模型。

随机误差是因为数据本身的随机所造成的误差，而非使用模型的误差。假如我们用一个很复杂的模型，可能将偏误降到很小，但是这并不是较那些一元而偏误高的模型更好。

因为复杂模型还是有随机误差存在，而且复杂模型花费的时间与成本更高，同时，复杂的数学并不见得被管理者所接受。

3.9.2 误差的衡量

在前面介绍时间数列法或焦点预测，我们不但预测未来值，同时在过去实际值也计算其预测值。误差的衡量皆是以预测值与实际值做比较。以下是几种衡量误差与公式。首先我们定义符号：

Y_i —— 第 i 期的实际值, $i = 1, \dots, n$;

F_i —— 第 i 期的预测值, $i = 1, \dots, n$;

n —— 已知 n 期的实际值与预测值。

1. 平均误差 (mean error, 简称 ME)

ME 是实际值与预测值之差, 然后加以平均, 即

$$ME = \sum_{i=1}^n (Y_i - F_i) / n$$

因为预测误差的数值 ($Y_i - F_i$) 可能有正值有负值, 相加结果会抵消误差, 所以 ME 可能会低估预测的误差。

2. 平均绝对差 (mean absolute deviation, 简称 MAD)

MAD 是实际值与预测值之差的绝对值, 然后加以平均, 即

$$MAD = \sum_{i=1}^n |Y_i - F_i| / n$$

MAD 是最常见的预测误差衡量, 因为我们可以从 MAD 计算预测的上下限, 以及计算追查信号 TS。当预测是正态分配时, MAD 与标准差之关系是

$$1 \text{ 个标准差} = \sqrt{\pi/2} \times MAD \approx 1.25MAD$$

所以利用统计方法, 预测值 ± 3 个标准差 (或 ± 3.75 MAD) 之内, 实际值应该有 99.7% 的概率会落在这个范围之内。

3. 均方误差 (mean squared error, 简称 MSE)

MSE 是实际值与预测值之差的平方和, 然后加以平均, 即

$$MSE = \sum_{i=1}^n (Y_i - F_i)^2 / n$$

虽然 MSE 类似统计的方差 (variance), 但是并不适合作为预测误差的衡量。因为 MSE 对误差是采取平方计算, 所以对于误差较大的, MSE 也较大。因此, MSE 只能作为一个参考的指标。

MSE 之运用和 MAD 相同, 不过计算标准差, 只要取 MSE 的平方根即可。

4. 平均绝对百分比误差 (mean absolute percentage error, 简称 MAPE)

MAD 或 MSE 的单位是和实际值与预测值相同, 所以对于数值大或变动大的时间数列, MAD 与 MSE 也相对很大, 比较没有客观的取舍标准, 所以利用 MAPE 可能较客观。

平均百分比误差 (mean percentage error, 简称 MPE), 其计算公式为

$$MPE = \sum_{i=1}^n (Y_i - F_i) / Y_i / n \times 100\%$$

$$MAPE = \sum_{i=1}^n |(Y_i - F_i) / Y_i| / n \times 100\%$$

5. 追查信号 (tracking signal, 简称 TS)

追查信号是利用 MAD 来计算, 其计算公式为

$$TS = \sum_{i=1}^n (Y_i - F_i) / MAD$$

若 TS 是正数，则表示预测值被低估；若 TS 是负数，则表示预测值被高估。TS 超过某个标准，则要追查预测方法是否要修改。至于 TS 的标准如何决定，则要视预测项目之种类（如物料管理 ABC 分类，A 类物料之价值高，则 TS 范围小，以便时常追查），也要看做计划或做预测的人员之时间（若没有时间追查，则 TS 范围大些）。

3.10 中文统计应用

3.10.1 简单移动平均法（例题 3.2）

执行“移动平均法”的操作示意图和结果分别如图 3-13 和图 3-14 所示。

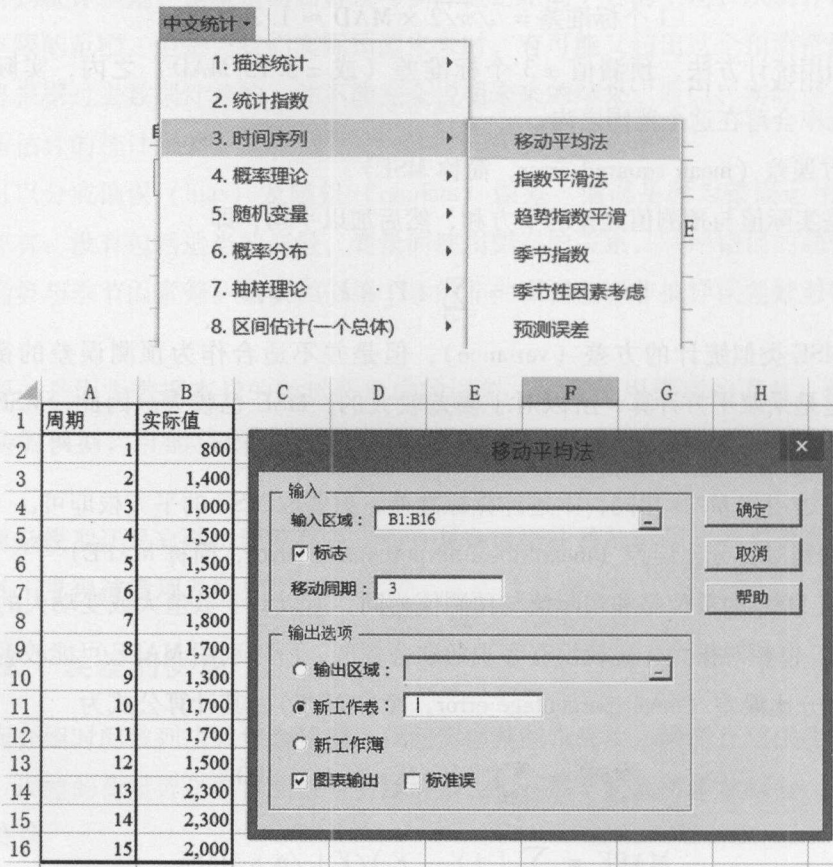


图 3-13 执行“移动平均法”的操作示意图

	A	B	C
1	简单移动平均		
2	时间	实际值	预测值
3		1	800
4		2	1400
5		3	1000
6		4	1500 1066.667
7		5	1500 1300
8		6	1300 1333.333
9		7	1800 1433.333
10		8	1700 1533.333
11		9	1300 1600
12		10	1700 1600
13		11	1700 1566.667
14		12	1500 1566.667
15		13	2300 1633.333
16		14	2300 1833.333
17		15	2000 2033.333
18		16	2200

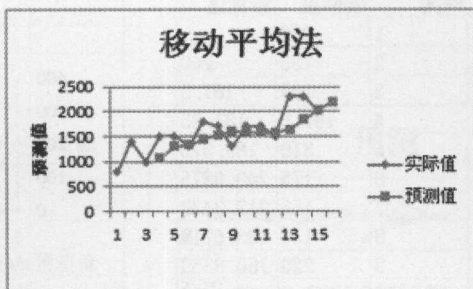


图 3-14 执行“移动平均法”的结果

3.10.2 指数平滑法 (例题 3.3)

执行“指数平滑法”的操作示意图和结果分别如图 3-15 和图 3-16 所示。

中文统计 -

- 1. 描述统计
- 2. 统计指数
- 3. 时间序列
- 4. 概率理论
- 5. 随机变量

- 移动平均法
- 指数平滑法
- 趋势指数平滑

	A	B
1	时间	实际值
2		1 200
3		2 135
4		3 195
5		4 197.5
6		5 310
7		6 175
8		7 155
9		8 130
10		9 220
11		10 277.5
12		11 235
13		

指数平滑法

输入

输入区域: B1:B12

平滑常数 (α): 0.5

☒ 标志

输出选项

☐ 输出区域:

☒ 新工作表:

☐ 新工作簿

☒ 图表输出 ☐ 标准误差

确定

取消

帮助

图 3-15 执行“指数平滑法”的操作示意图

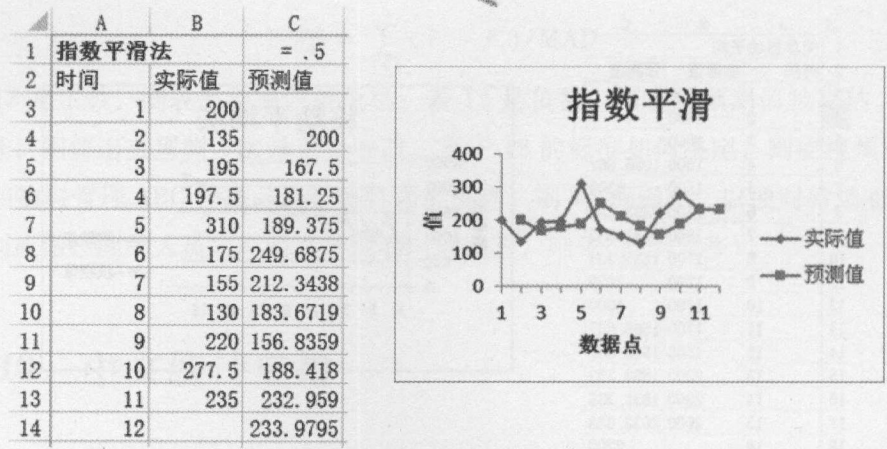


图 3-16 执行“指数平滑法”的结果

3. 10.3 趋势指数平滑（例题 3.6）

执行“趋势指数平滑”的操作示意图和结果分别如图 3-17 和图 3-18 所示。

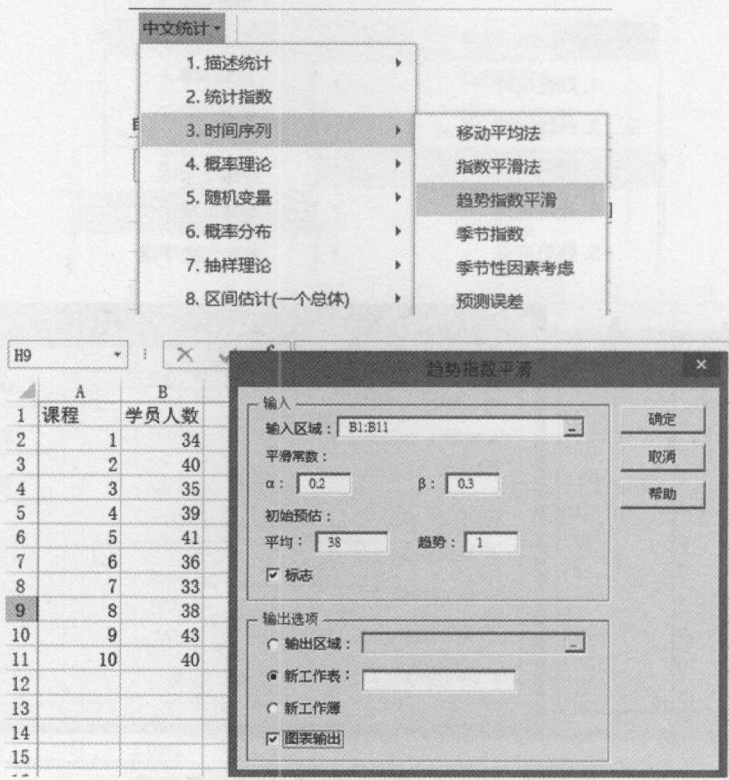


图 3-17 执行“趋势指数平滑”的操作示意图

	A	B	C	D	E	F
1	趋势指数平滑法					
2						
3	Smoothing constants					
4	Alpha	0.2				
5	Beta	0.3				
6						
7	初始预估					
8	平均	38				
9	趋势	1				
10						
11	期间	学员人数	LastTrend	EstiTrend	预测值	
12	1	34	0	1	39	
13	2	40	0.96	0.7	38.7	
14	3	35	-0.1696	0.778	39.738	
15	4	39	0.436896	0.49372	39.28412	
16	5	41	0.735879	0.476673	39.70397	
17	6	36	-0.34909	0.554435	40.51761	
18	7	33	-1.09612	0.283378	39.89747	
19	8	38	-0.20797	-0.13047	38.3875	
20	9	43	0.815024	-0.15372	38.15628	
21	10	40	0.284517	0.136903	39.26193	
22	11		-7.73696	0.181187	39.59073	

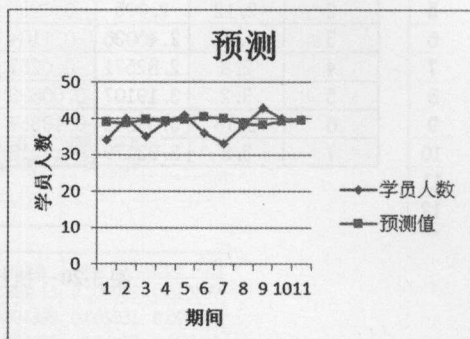


图 3-18 执行“趋势指数平滑”的结果

3.10.4 一元线性回归（例题 3.5）

执行“一元线性回归”的操作示意图和结果分别如图 3-19 和图 3-20 所示。

12. 回归与相关

13. 分类数据分析

一元线性回归

一元与多元线性回归

	A	B
1	t	S
2	1	1.76
3	2	2.12
4	3	2.35
5	4	2.8
6	5	3.2
7	6	3.75
8	7	3.8

一元线性回归

输入

输入Y区域:

输入X区域:

置信度: (X,Y输入区域不含标志)

☐ 检验相关系数

☐ 检验 β_0

图 3-19 执行“一元线性回归”的操作示意图

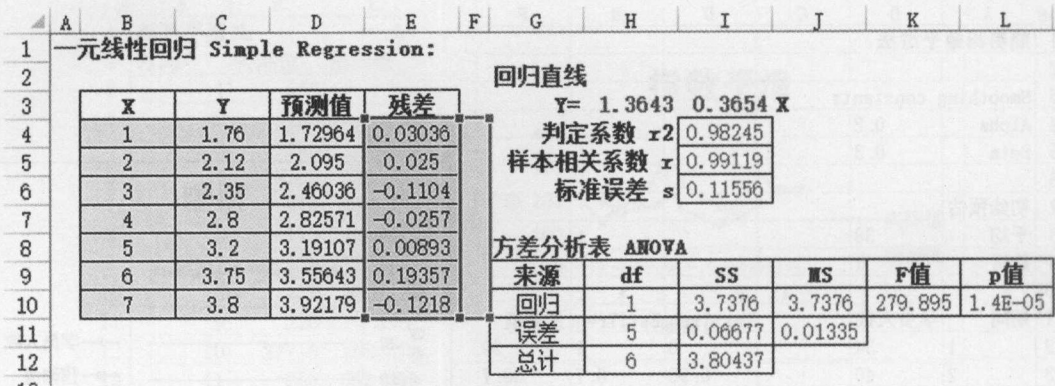


图 3-20 执行“一元线性回归”的结果

3.10.5 一元与多元线性回归（例题 3.10）

执行“一元与多元线性回归”的操作示意图和结果分别如图 3-21 和图 3-22 所示。



图 3-21 执行“一元与多元线性回归”的操作示意图

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	回归统计						
4	Multiple R	0.943022					
5	R Square	0.88929					
6	Adjusted R	0.859767					
7	标准误差	0.03806					
8	观测值	20					
9							
10	方差分析						
11		df	SS	MS	F	Significance F	
12	回归分析	4	0.174531	0.043633	30.12217	5.2E-07	
13	残差	15	0.021728	0.001449			
14	总计	19	0.196259				
15							
16		Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	0.559588	0.02374	23.57193	2.88E-13	0.508988	0.610187
18	t	0.005038	0.001504	3.348434	0.004399	0.001831	0.008244
19	Q1	0.003513	0.02449	0.143423	0.887865	-0.04869	0.055713
20	Q2	0.139275	0.024258	5.74134	3.9E-05	0.08757	0.19098
21	Q3	0.205238	0.024118	8.509752	3.99E-07	0.153831	0.256644

图 3-22 执行“一元与多元线性回归”的结果

3.10.6 季节指数 (例题 3.7, 例题 3.8)

执行“季节指数”的操作示意图和结果如图 3-23 所示。

	A	B
1	1	0.561
2	2	0.702
3	3	0.8
4	4	0.568
5	1	0.575
6	2	0.738
7	3	0.868
8	4	0.605
9	1	0.594
10	2	0.738
11	3	0.729
12	4	0.6
13	1	0.622
14	2	0.708
15	3	0.806
16	4	0.632
17	1	0.665
18	2	0.835
19	3	0.873
20	4	0.67

	A	B
1	季节指数	
2	季节	季节指数
3	1	0.8813
4	2	1.0747
5	3	1.1726
6	4	0.8713

图 3-23 执行“季节指数”的操作示意图和结果

3.11 本章流程图

本章的流程图如图 3-24 所示。

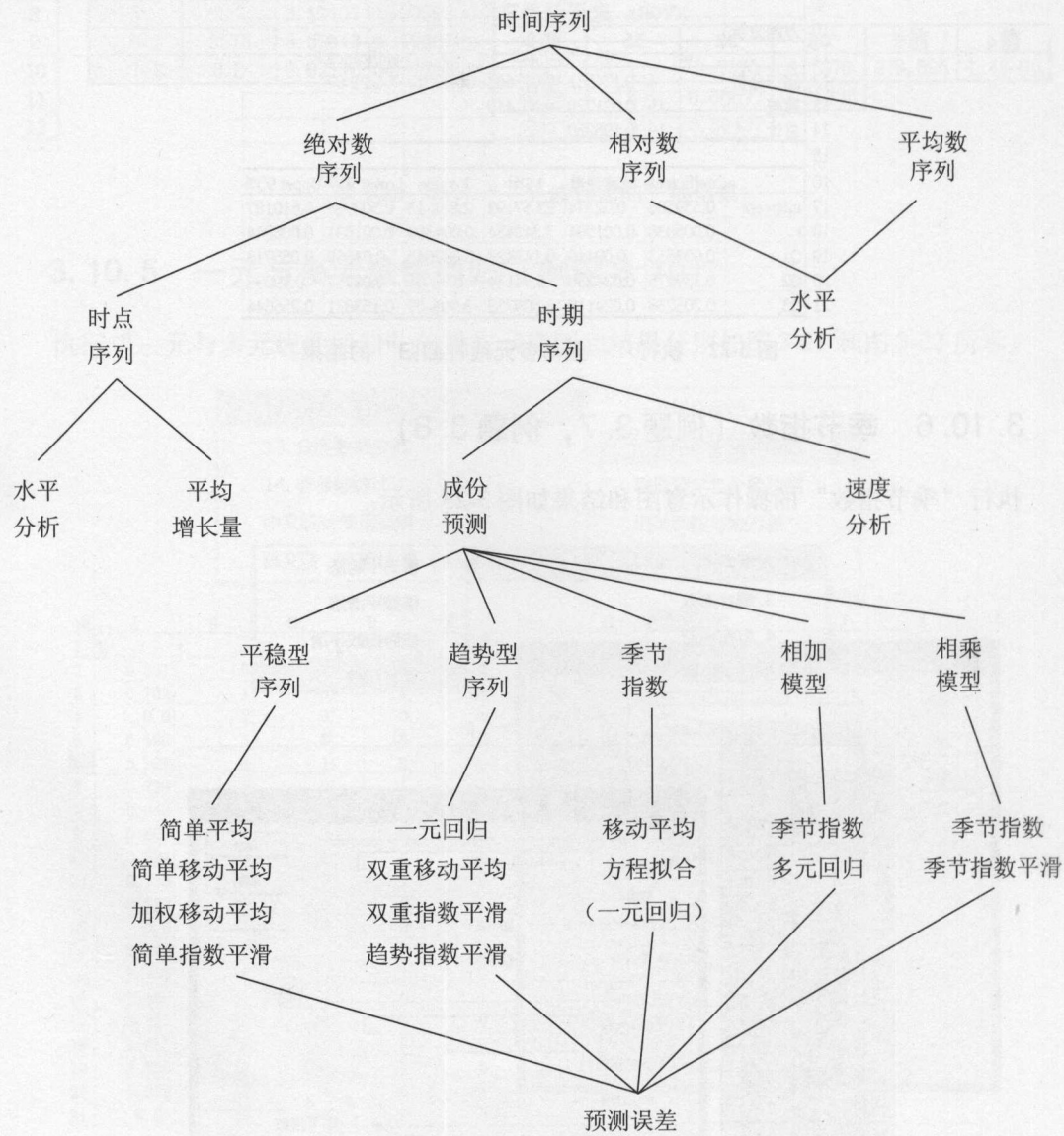


图 3-24 第 3 章的流程图

3.12 本章思维导图

本章思维导图如图 3-25 所示。

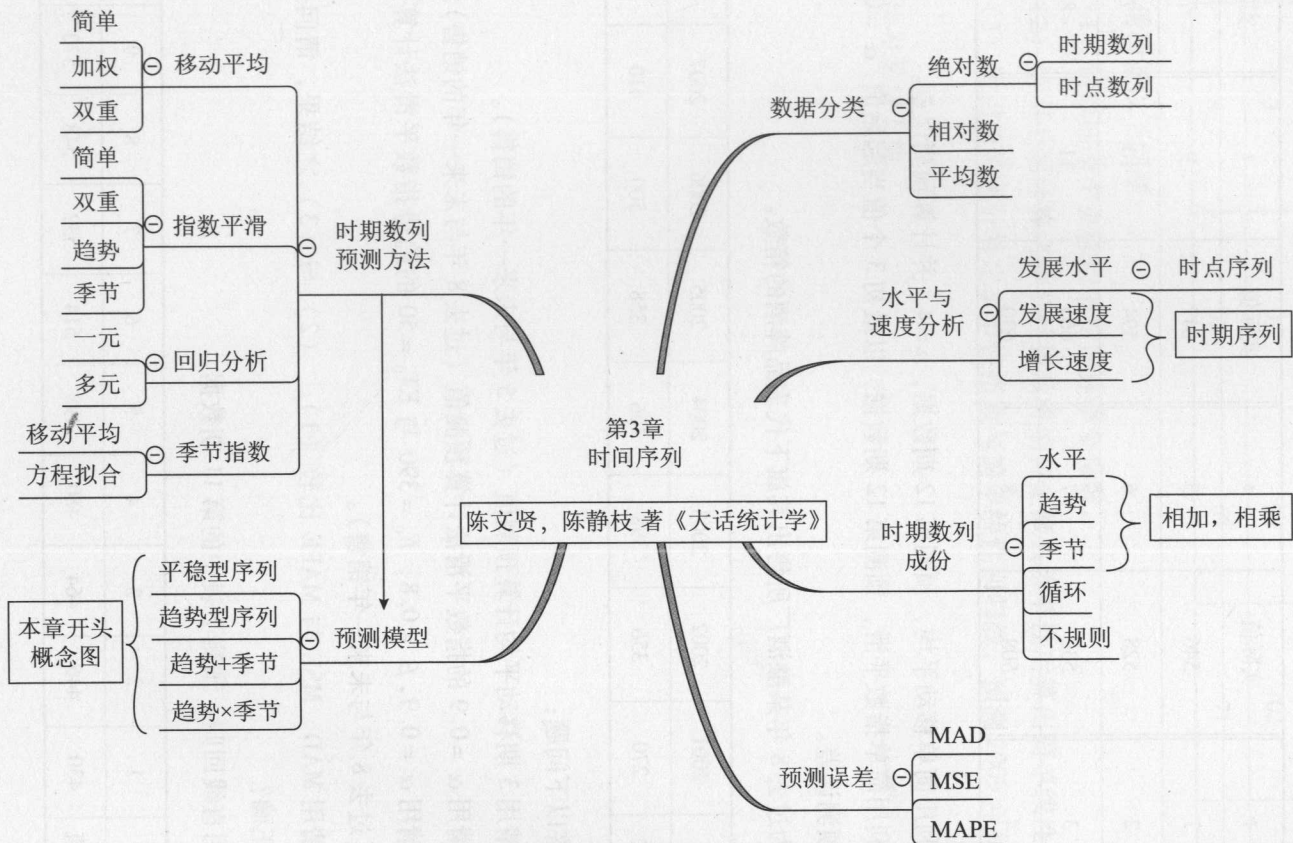


图3-25 第3章思维思维

习 题

1. 下列时间数列数据：

t	实际值	t	实际值	t	实际值
1	546	5	647	9	736
2	528	6	594	10	724
3	530	7	665	11	813
4	508	8	630	12	

- (1) 利用简单移动平均，预测第 12 期数据， $n = 4$ ，并计算预测误差。
- (2) 利用简单指数平滑，预测第 12 期数据，以最初 3 个值当起始值， $\alpha = 0.1$ ，并计算预测误差。
2. 下表为过去 8 年某酿酒厂所售出其旗下代表品牌酒的箱数。

年	2001	2002	2003	2004	2005	2006	2007	2008
箱数	270	356	398	456	358	500	410	376

- 请回答以下问题：
- (1) 请用 3 期移动平均计算预测值（过去 5 年与未来一年的销售）。
- (2) 请用 $\alpha = 0.9$ 的指数平滑法计算预测值（过去 8 年与未来一年的销售）。
- (3) 请用 $\alpha = 0.9, \beta = 0.8, F_0 = 390$ 与 $ET_0 = 50$ 的趋势指数平滑法计算预测值（过去 8 年与未来一年销售）。
- (4) 请用 MAD、MSE 与 MAPE 比较 (1)、(2) 与 (3) 之结果，请问何者较正确？
3. 试利用直线回归，预测下表中的第 11 期数据。

时间	1	2	3	4	5	6	7	8	9	10
实际值	430	446	464	480	498	514	532	548	570	591

4. 下表为某个人计算机供货商从 2005 至 2008 年的每季销售计算机金额（以百万元计）。

季 \ 年	2005	2006	2007	2008
1	60	65	68	74
2	75	83	85	90
3	93	98	102	106
4	62	69	71	75

请回答以下问题：

- (1) 请以四季中央移动平均计算季节指数。
- (2) 请用上题计算之季节性指数将销售金额去季节性后算出简单线性回归线。
- (3) 请用前两题计算之季节性指数与简单线性回归线，预测 2009 年每季之销售。

其他习题请下载。



第4章

统计指数

物有本末，事有始终，知所先后，则近道矣。

——《大学》

求则得之，舍则失之，是求有益于得也，求在我者也。

——《孟子·尽心篇》

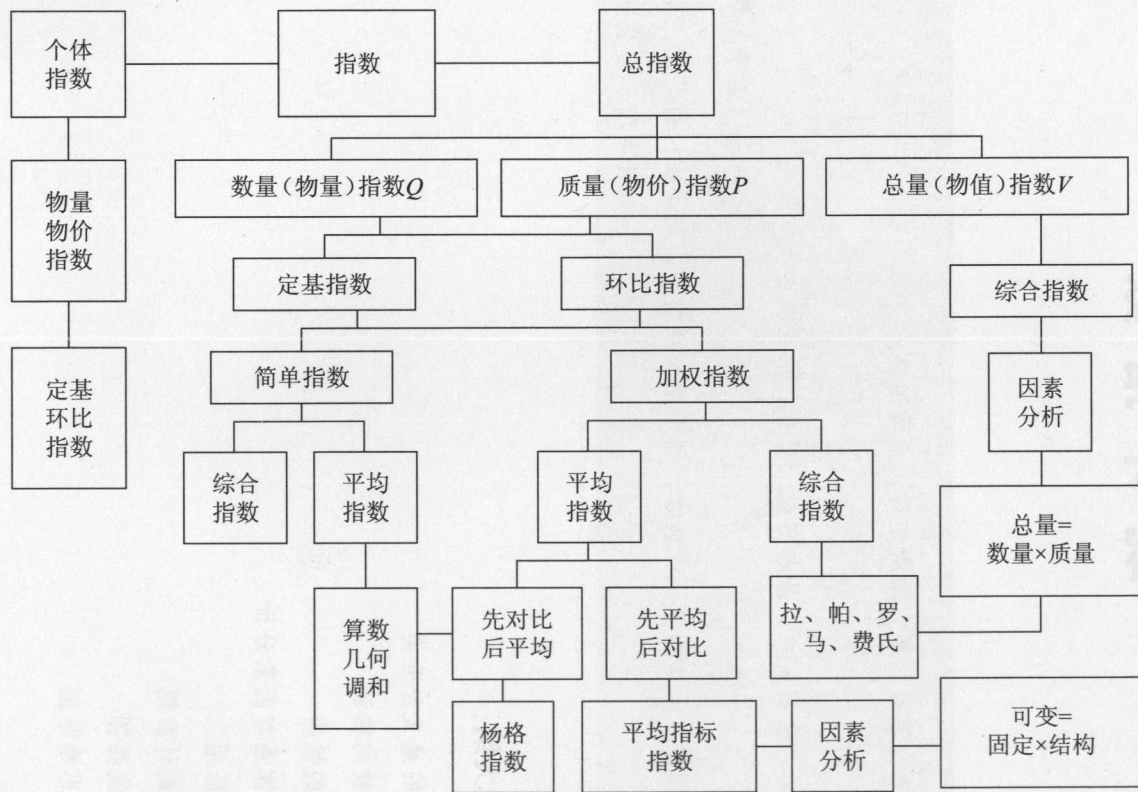
凡事豫则立，不豫则废。言前言，则不殆；事前定，则不困；行前定，则不疚。

——《中庸·哀公问政篇》



本章重点大纲：

- 4.1 指数的意义与分类
- 4.2 总指数的编制
- 4.3 指数的性质
- 4.4 指数体系与因素分析
- 4.5 指数应用
- 4.6 中文统计应用
- 4.7 本章流程图
- 4.8 本章思维导图



本章概念图

4.1 指数的意义与分类

指数 (index) 的主要功能是计算某一经济活动, 在不同时间 (例如: 物价指数、股价指数), 或不同地点 (例如: 生活指数、生产量指数), 做比较而得到的相对数值, 以表示变动的程度。

指数表示不同时间 (或同时间不同地点) 的经济活动的变化。因为许多工商数据的计量单位不同 (食物就有不同的单位, 还有进出口的产品等), 不能加以比较。指数的另一个功能是, 综合不同单位的数据, 加以统计比较。例如: 趸售物价指数 (wholesale price index), 工业生产指数 (industrial production index) 等。

物价指数是最常用的指数, 它显示货币的购买力, 可以用以计算币值的变动, 表示实质所得, 应用于薪资调整, 企业的资产重估等方面。

比较两个数量, 必须以某一个数量为基础, 此基础数量称作基数 (基期指数 base), 基数所在的时间称作基期 (base period)。报告期 (given period) 是要计算指数的时期。

从指数数列 (不同时间对基期的指数), 有时可以看出经济活动的长期趋势, 季节变动或循环变化, 请见图 4-1。

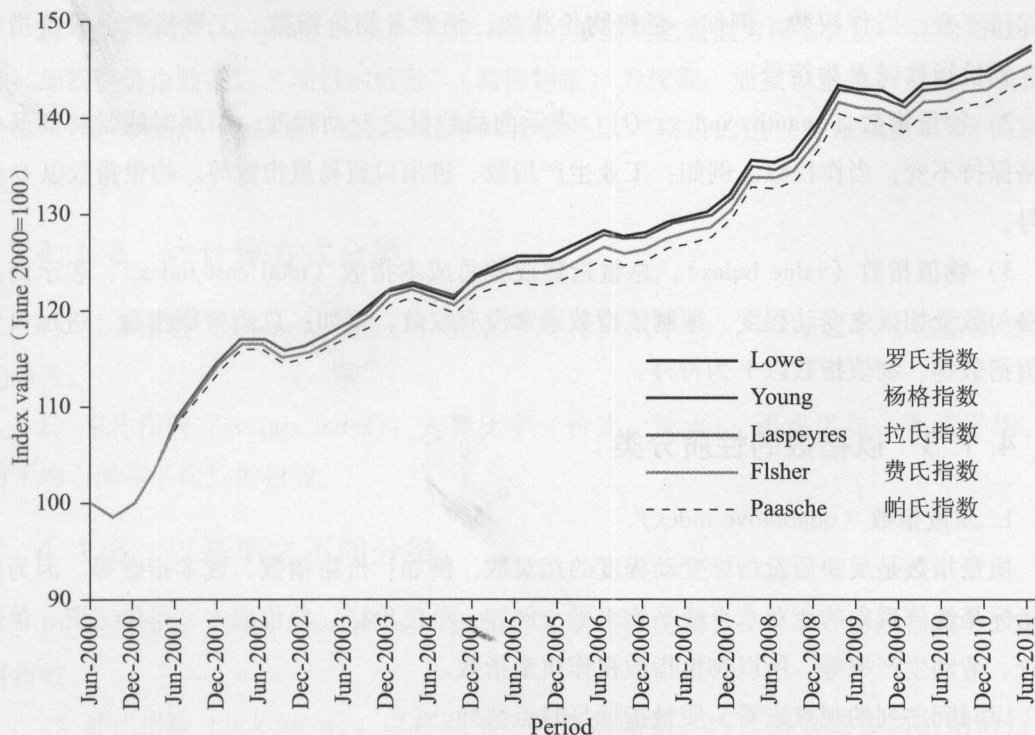


图 4-1 指数的比较 (拉氏 \geq 费氏 \geq 帕氏, 罗氏和杨格固定权数) (参考英文书目 [2])

所以,指数有下列性质。

- 1) 综合性:指数通常以多种项目做综合、平均或加权计算,所以有综合变动的结果。
 - 2) 代表性:指数并非以全部项目做计算,而是挑选重要的部分项目作代表计算。
 - 3) 相对性:指数所表示的变动,是相对的变动。
 - 4) 平均性:指数通常是以平均方法产生,其变动是多种项目(商品)的平均变动。
- 指标是数量、价格、平均指标等,是有单位的。指数是相对数,是没有单位的。
- 指数的分类,以不同的方式分类,有下列6种分类。

4.1.1 以商品的数目分类

- 1) 个体指数(individual index):表示单项事物或商品之变动的相对数,例如:一种产品的个体产量指数,或一种商品的个体价格指数。
- 2) 总指数(combined index):表示多项事物或商品之变动的相对数,例如:零售物价总指数、工业产品产量总指数、商品销售量总指数。编制时通常将商品的价格保持不变,当作权数。本章的统计指数主要是以总指数。

4.1.2 以指数的用途分类

- 1) 物价指数(price index, P):表示商品价格之变动程度,编制时通常将商品的数量保持不变,当作权数。例如:趸售物价指数、消费者物价指数、工资指数、股价指数等。物价指数以 P 为符号。
- 2) 物量指数(quantity index, Q):表示商品数量之变动程度,编制时通常将商品的价格保持不变,当作权数。例如:工业生产指数、进出口贸易量指数等。物量指数以 Q 为符号。
- 3) 物值指数(value index),总量指数或称总成本指数(total cost index):表示商品价格与数量相乘之变动程度,编制该指数通常没有权数。例如:总销售额指数、进出口贸易值指数等。物值指数以 V 为符号。

4.1.3 以指数的性质分类

1. 质量指数(qualitative index)

质量指数是反映质量指标变动程度的相对数,例如:价格指数、成本指数等。因为质量指标是经济现象的相对水平或平均水平,例如:单位价格、单位成本、单位工资、单位股价、劳动生产率等,所以物价指数视作质量指数。

以时间序列的观点来看,质量指标是时点数列。

2. 数量指数 (quantitative index)

数量指数是反映数量指标变动程度的相对数,例如:产量指数、销售量指数等。数量指标通常采用商品实物的计量的单位,配合质量指标计价的单位。

以时间序列的观点来看,数量指标是时期数列。

3. 总量指数 (total quantity index) 或称总成本指数 (total cost index)

总量指数是反映总金额或总成本指标变动程度的相对数,表示商品价格与数量相乘之变动程度,例如:销售额指数、营业额指数等。以时间序列的观点来看,总量指标是时期数列。

$$\text{总量} = \text{数量} \times \text{质量}$$

数量指标和质量指标的区别是相对的,例如:

$$\text{原材料支出总额} = \text{产量} \times \text{单耗} \times \text{单位原材料价格}$$

“单耗”是每单位产量消耗原材料的单位,单耗指标对产量指标来说是质量指标,单耗指标对单位原材料价格来说是数量指标。所以本章指数,较常用物价指数和物量指数来称呼。

4.1.4 以有无加权分类

1) 简单指数 (simple index): 没有用权数计算的指数。

2) 加权指数 (weighted index): 以各项目的重要性之数值为权数,所计算的指数。例如:加权物价指数是以“项目的数量”(简称物量)为权数;加权物量指数是以“项目的价格”(简称物价)为权数;加权价比平均指数是以“项目的值(价乘以量)”(简称物值)为权数。

4.1.5 以计算方式分类

1) 综合指数 (aggregate index): 先求平均(算术平均),再算比率(与基期相比)的指数。

2) 平均指数 (average index): 先算比率(价比、量比),再求平均(算术平均、几何平均、调和平均)的指数。

4.1.6 以基期之不同分类

1) 定基指数 (fixed base index): 以固定基期计算的指数。4.2节的总指数公式是定基指数。

2) 环比指数 (link index): 以移动基期计算的指数。环比指数是移动基期的指数,

以前一期为基期，编算本期的指数。环比指数的计算公式与定基指数相同，只是每期选用的商品项目可能不同。例如： $I_{(t-1)t}$ 是以 $t-1$ 期为基期，报告期为 t 期的环比指数； $I_{t(t+1)}$ 是以 t 期为基期，报告期为 $t+1$ 期的环比指数。但是 $I_{(t-1)t}$ 与 $I_{t(t+1)}$ 计算时所采用的商品项目可能不同。 $I_{(t-1)t}$ 是将定基指数公式中的 p_0, q_0 改为 p_{t-1}, q_{t-1} ，将 p_1, q_1 改为 p_t, q_t ，例如：环比拉氏物价指数的公式为

$$P_{L,(t-1)t} = \frac{\sum p_t q_{t-1}}{\sum p_{t-1} q_{t-1}}$$

3) 链指数或锁指数 (chain index)：环比指数相乘而得的定基指数。链指数是连续期数的环比指数相乘，以 C_{0t} 表示。链指数和定基指数不同的，是考虑每期不同商品和权数。其计算公式为

$$C_{0t} = I_{01} I_{12} I_{23} \cdots I_{(t-1)t}$$

另外指数分类有动静态指数和静态指数（空间指数），本章只讨论前者。

个体指数与总体指数的比较如表 4-1 所示。

表 4-1 个体指数与总指数的比较

指数名称		输入数据 (input)			
		基期	报告期	商品数目	数据
个体指数		1 个	n 个	1 个	一个变量：数量或价格（质量）时间序列→发展速度
总指数	简单指数	1 个	1 个	多个	一个变量：数量或价格（质量）
	加权指数	1 个	1 个	多个	两个变量：数量和价格（质量）

总指数的报告期，在本书及中文统计，只有一个报告期，实际应用当然扩大到 n 个报告期。

个体指数的公式（个体指数只有定基指数环比指数，没有简单指数与加权指数，当然没有综合指数与平均指数）如下

定基物价指数 $I = \frac{p_1}{p_0}$ ；定基物量指数 $I = \frac{q_1}{q_0}$

环比物价指数 $I = \frac{p_i}{p_{i-1}}$ ；环比物量指数 $I = \frac{q_i}{q_{i-1}}$

式中： p_0 ——商品在基期的价格；

p_1 ——商品在报告期的价格；

q_0 ——商品在基期的数量；

q_1 ——商品在报告期的数量。

例题 4.1 假设以 3 种食品为代表项目，基期是 2010 年，当期是 2015 年，食品的价格及数量如表 4-2 所示。计算各种指数。解答如 4.2 节。

表 4-2 商品物价及数量

商品名称	数量		单价		权数				个体指数	
	基期 q_0	报告期 q_1	基期 p_0	报告期 p_1	q_0p_0	q_1p_1	q_1p_0	q_0p_1	量比 q_1/q_0	价比 p_1/p_0
甲	50	40	\$ 22	\$ 30	1100	1200	880	1500	0.8	1.36
乙	2	3	\$ 20	\$ 20	40	60	60	40	1.5	1.0
丙	80	100	\$ 5	\$ 6	400	600	500	480	1.25	1.2
Σ	132	143	\$ 47	\$ 56	1540	1860	1440	2020	3.55	3.56

4.2 总指数的编制

以下各个总指数公式，是利用表 4-2 的数据计算。总和 Σ 是针对商品的数目加总。请注意，计算公式都有“对比”，（所以才叫指数），只是和“综合或平均”有先后的关系。

4.2.1 简单综合指数

$$\text{物价指数 } I_P = \frac{\sum_{i=1}^n \frac{p_{1i}}{n}}{\sum_{i=1}^n \frac{p_{0i}}{n}} = \frac{\sum_{i=1}^n p_{1i}}{\sum_{i=1}^n p_{0i}} = \frac{\sum p_1}{\sum p_0} = \frac{56}{47} = 1.10$$

$$\text{物量指数 } I_Q = \frac{\sum_{i=1}^n \frac{q_{1i}}{n}}{\sum_{i=1}^n \frac{q_{0i}}{n}} = \frac{\sum_{i=1}^n q_{1i}}{\sum_{i=1}^n q_{0i}} = \frac{\sum q_1}{\sum q_0} = \frac{143}{132} = 1.08$$

式中： n ——商品的数目；

q_{0i} ——第 i 商品在基期的数量；

q_{1i} ——第 i 商品在报告期的数量；

p_{0i} ——第 i 商品在基期的价格；

p_{1i} ——第 i 商品在报告期的价格。

简单综合指数有时没有意义，因为商品的单位不同，价格相加没有意义。

4.2.2 简单平均指数

简单算术平均指数：

$$\text{物价指数 } A = \frac{\sum_{i=1}^n \frac{p_{1i}}{p_{0i}}}{n} = \frac{\sum \frac{p_1}{p_0}}{n} = \frac{3.56}{3} = 1.19$$

$$\text{物量指数 } A = \frac{\sum_{i=1}^n \frac{q_{1i}}{q_{0i}}}{n} = \frac{\sum \frac{q_1}{q_0}}{n} = \frac{3.55}{3} = 1.18$$

简单几何平均指数：

$$\text{物价指数 } G = \sqrt[n]{\prod_{i=1}^n \left(\frac{p_{1i}}{p_{0i}} \right)} = \sqrt[n]{\prod \frac{p_1}{p_0}} = \sqrt[3]{1.36 \times 1 \times 1.2} = 1.18$$

$$\text{物量指数 } G = \sqrt[n]{\prod_{i=1}^n \left(\frac{q_{1i}}{q_{0i}} \right)} = \sqrt[n]{\prod \frac{q_1}{q_0}} = \sqrt[3]{0.8 \times 1.5 \times 1.25} = 1.14$$

简单调和平均指数：

$$\text{物价指数 } H = \frac{n}{\sum_{i=1}^n \frac{p_{0i}}{p_{1i}}} = \frac{n}{\sum \frac{p_0}{p_1}} = \frac{3}{\frac{1}{1.36} + \frac{1}{1} + \frac{1}{1.2}} = 1.17$$

$$\text{物量指数 } H = \frac{n}{\sum_{i=1}^n \frac{q_{0i}}{q_{1i}}} = \frac{n}{\sum \frac{q_0}{q_1}} = \frac{3}{\frac{50}{40} + \frac{2}{3} + \frac{80}{100}} = 1.1$$

4.2.3 加权综合指数

加权综合指数（物价或物量）有 6 种公式，分别采用不同的时期的“数量”或“单价”作权数，分子和分母的单位是“数量 × 单价”（ $q \times p$ ）。

1) 拉氏指数（Laspeyres index）：以基期的数量 q_0 为权数。

$$\text{物价指数 } P_L = \frac{\sum p_1 q_0}{\sum p_0 q_0} = \frac{2020}{1540} = 1.31$$

$$\text{物量指数 } Q_L = \frac{\sum p_0 q_1}{\sum p_0 q_0} = \frac{1440}{1540} = 0.94$$

2) 帕氏指数（Paasche index）：以报告期的数量 q_1 或单价 p_1 为权数。

$$\text{物价指数 } P_P = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{1860}{1440} = 1.29$$

$$\text{物量指数 } Q_p = \frac{\sum p_1 q_1}{\sum p_1 q_0} = \frac{1860}{2020} = 0.92$$

3) 罗氏指数 (Lowe index): 以基期及报告期以外的固定期 (a 期) 的数量 q_a 或单价 p_a 为权数。假设 3 个商品 a 期的数量 q_a 分别是 45, 2.5, 90, 单价 p_a 分别是 \$25, \$20, \$5.5。

$$\text{物价指数 } P_{Lo} = \frac{\sum p_1 q_a}{\sum p_0 q_a} = \frac{30 \times 45 + 20 \times 2.5 + 6 \times 90}{22 \times 45 + 20 \times 2.5 + 5 \times 90} = \frac{1940}{1490} = 1.30$$

$$\text{物量指数 } Q_{Lo} = \frac{\sum p_a q_1}{\sum p_a q_0} = \frac{40 \times 25 + 3 \times 20 + 100 \times 5.5}{50 \times 25 + 2 \times 20 + 80 \times 5.5} = \frac{1610}{1730} = 0.93$$

4) 马氏指数或马埃指数 (Marshall - Edgeworth index): 以基期及报告期的数量的平均 $(q_0 + q_1)/2$ 为权数。

$$\text{物价指数 } P_M = \frac{\sum p_1 (q_0 + q_1)/2}{\sum p_0 (q_0 + q_1)/2} = \frac{\sum p_1 (q_0 + q_1)}{\sum p_0 (q_0 + q_1)} = 1.30$$

$$\text{物量指数 } Q_M = \frac{\sum q_1 (p_0 + p_1)/2}{\sum q_0 (p_0 + p_1)/2} = \frac{\sum q_1 (p_0 + p_1)}{\sum q_0 (p_0 + p_1)} = 0.93$$

5) 费氏指数或称费雪指数 (Fisher index): 以拉氏及帕氏两个指数, 取几何平均, 又称理想指数 (ideal index)。

$$\text{物价指数 } P_F = \sqrt{P_L \times P_P} = \sqrt{1.31 \times 1.29} = 1.30$$

$$\text{物量指数 } Q_F = \sqrt{Q_L \times Q_P} = \sqrt{0.94 \times 0.92} = 0.93$$

6) 华氏指数 (Walsh index): 以基期及报告期的数量的几何平均 $\sqrt{q_0 q_1}$ 为权数。

$$\text{物价指数 } P_w = \frac{\sum p_1 \sqrt{q_0 q_1}}{\sum p_0 \sqrt{q_0 q_1}} = \frac{30 \sqrt{50 \times 40} + 20 \sqrt{2 \times 3} + 6 \sqrt{80 \times 100}}{22 \sqrt{50 \times 40} + 20 \sqrt{2 \times 3} + 5 \sqrt{80 \times 100}} = \frac{1927.24}{1480.04} = 1.30$$

$$\text{物量指数 } Q_w = \frac{\sum q_1 \sqrt{p_0 p_1}}{\sum q_0 \sqrt{p_0 p_1}} = \frac{40 \sqrt{22 \times 30} + 3 \sqrt{20 \times 20} + 100 \sqrt{5 \times 6}}{50 \sqrt{22 \times 30} + 2 \sqrt{20 \times 20} + 80 \sqrt{5 \times 6}} = \frac{1635.30}{1762.66} = 0.93$$

4.2.4 加权平均指数

加权价比平均物价指数, 因为价比或量比是没有单位的数值 (物价相除), 而分子和分母的单位是“数量 \times 单价” ($q \times p$), 所以要用物值 $q \times p$ 作权数。

加权价比平均物价指数有 12 个公式, 因为有 3 种平均方法 (算术、几何、调和), 每种平均方法又有 4 种权数 (值权): $p_0 q_0$ 、 $p_0 q_1$ 、 $p_1 q_0$ 、 $p_1 q_1$ 。

1. 加权平均指数：以固定期 $p_a q_a$ 当权数

$$\text{物价指数 } P_A = \frac{\sum \left(\frac{p_1}{p_0}\right) p_0 q_0}{\sum p_0 q_0} = P_L$$

$$\text{物量指数 } Q_A = \frac{\sum \left(\frac{q_1}{q_0}\right) p_0 q_0}{\sum p_0 q_0} = Q_L$$

除了上述权数，还可以选其他时期的物值 $p_a q_a$ 作权数；选 $(p_0 q_0 + p_1 q_1)/2$ 作权数；或选几何平均物值 $\sqrt{p_0 q_0 \times p_1 q_1}$ 作权数等。

2. 杨格指数 (Young index)：以固定期 $p_a q_a$ 当权数

假设 3 个商品 a 期的数量 q_a 分别是 45, 2.5, 90, 单价 p_a 分别是 \$25, \$20, \$5.5。

$$\text{物价指数 } P_Y = \frac{\sum \left(\frac{p_1}{p_0}\right) p_a q_a}{\sum p_a q_a} = \frac{1.36 \times 45 \times 25 + 1 \times 2.5 \times 20 + 1.2 \times 90 \times 5.5}{45 \times 25 + 2.5 \times 20 + 90 \times 5.5} = \frac{2174}{1670} = 1.3$$

$$\text{物量指数 } Q_Y = \frac{\sum \left(\frac{q_1}{q_0}\right) p_a q_a}{\sum p_a q_a} = \frac{0.8 \times 45 \times 25 + 1.5 \times 2.5 \times 20 + 1.25 \times 90 \times 5.5}{45 \times 25 + 2.5 \times 20 + 90 \times 5.5} = 0.95$$

请注意，杨格指数和国内多数统计学书本的公式定义不同，有一些书本将“罗氏指数”的公式定义为“杨格指数”。请参考本书英文参考著作 [2]、[4]、[11]、[12]。

3. 加权几何平均指数：以固定期 $p_0 q_0$ 当权数

$$P_G = \sqrt[\sum p_0 q_0]{\prod \left(\frac{p_1}{p_0}\right)^{p_0 q_0}}$$

$$\log P_G = \frac{\sum [p_0 q_0 (\log p_1 - \log p_0)]}{\sum p_0 q_0}$$

说明，G 为加权几何平均指数。

4.2.5 物值指数

物值指数没有加权指数，只有简单指数，简单综合指数的公式是：

物值指数 = 当期物值总和 ÷ 基期物值总和

$$V = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{1860}{1540} = 1.21$$

例题 4.2 总指数的计算。(解答见网络资源)

总指数（质量指数）的分类表如表 4-3 所示。

表 4-3 总指数（质量指数）的分类表

分类 1	分类 2	分类 3	分类 4	公式	计算方法/指数名称
物价指数 质量指数	简单指数	综合指数		$I_p = \frac{\sum p_1}{\sum p_0}$	先综合（不加权）， 后对比 简单综合物价指数
		平均指数	算术平均	$A = \frac{\sum \frac{p_1}{p_0}}{n}$	先对比（个体指数）， 后算术平均（不加权） 简单算术平均物价指数
			几何平均	$G = \sqrt[n]{\prod \left(\frac{p_1}{p_0}\right)}$	先对比（个体指数）， 后几何平均（不加权） 简单几何平均物价指数
			调和平均	$H = \frac{n}{\sum \frac{p_0}{p_1}}$	先对比（注意倒数）， 后调和平均（不加权） 简单调和平均物价指数
物价指数 质量指数	加权指数	综合指数	以基期数量加权	$P_L = \frac{\sum p_1 q_0}{\sum p_0 q_0}$	先综合（加权）， 后对比 拉氏物价指数
			以报告期数量加权	$P_P = \frac{\sum p_1 q_1}{\sum p_0 q_1}$	先综合（加权）， 后对比 帕氏物价指数
			以两期数量加权	$P_M = \frac{\sum p_1 (q_0 + q_1)}{\sum p_0 (q_0 + q_1)}$	先综合（加权）， 后对比 马氏物价指数
			以固定 期数量 加权	$P_{Lo} = \frac{\sum p_1 q_a}{\sum p_0 q_a}$	先综合（加权）， 后对比 罗氏物价指数
			以拉氏 帕氏几何平均	$P_F = \sqrt{P_L \times P_P}$	费氏物价指数
		平均指数	算术平均 以 pq 加权	$P_A = \frac{\sum \left(\frac{p_1}{p_0}\right) p_0 q_0}{\sum p_0 q_0} = P_L$	先对比（个体指数）， 后平均（加权算术） 加权算术平均物价指数
				$P_Y = \frac{\sum \left(\frac{p_1}{p_0}\right) p_a q_a}{\sum p_a q_a}$	先对比（个体指数）， 后平均（加权算术） 杨格物价指数
			调和平均	$P_H = \frac{\sum p_0 q_0}{\sum \left(\frac{p_0}{p_1}\right) p_0 q_0}$	先对比（个体指数）， 后平均（加权调和） 加权调和平均物价指数
			算术平均	$\frac{\bar{p}_1}{\bar{p}_0} = \frac{\sum p_1 q_1 / \sum q_1}{\sum p_0 q_0 / \sum q_0}$	先平均（加权）， 后对比 平均指标物价指数

总指数（数量指数）的分类表如表 4-4 所示。

表 4-4 总指数（数量指数）的分类表

分类 1	分类 2	分类 3	分类 4	公式	计算方法/指数名称	
数量指数	简单指数	综合指数		$I_Q = \frac{\sum q_1}{\sum q_0}$	先综合（不加权）， 后对比 简单综合物量指数	
		平均指数	算术平均	$A = \frac{\sum \frac{q_1}{q_0}}{n}$	先对比（个体指数）， 后算术平均（不加权） 简单算术平均物量指数	
			几何平均	$G = \sqrt[n]{\prod \left(\frac{q_1}{q_0}\right)}$	先对比（个体指数）， 后几何平均（不加权） 简单几何平均物量指数	
			调和平均	$H = \frac{n}{\sum \frac{q_0}{q_1}}$	先对比（注意倒数）， 后调和平均（不加权） 简单调和平均物量指数	
	加权指数	综合指数 以物价 P 加权	以基期价格加权	$Q_L = \frac{\sum q_1 p_0}{\sum q_0 p_0}$	先综合（加权）， 后对比 拉氏物量指数	
			以报告期价格加权	$Q_P = \frac{\sum q_1 p_1}{\sum q_0 p_1}$	先综合（加权）， 后对比 帕氏物量指数	
			以两期价格加权	$Q_M = \frac{\sum q_1 (p_0 + p_1)}{\sum q_0 (p_0 + p_1)}$	先综合（加权）， 后对比 马氏物量指数	
			以固定价格加权	$Q_{Lo} = \frac{\sum q_1 p_a}{\sum q_0 p_a}$	先综合（加权）， 后对比 罗氏物量指数	
			以拉氏、帕氏几何平均	$Q_F = \sqrt{Q_L \times Q_P}$	费氏物量指数	
		平均指数	以 pq 加权	$Q_A = \frac{\sum \left(\frac{q_1}{q_0}\right) p_0 q_0}{\sum p_0 q_0} = Q_L$	先对比（个体指数）， 后平均（加权） 加权算术平均物量指数	
			以 p 加权	$\frac{\bar{q}_1}{\bar{q}_0} = \frac{\sum p_1 q_1 / \sum p_1}{\sum p_0 q_0 / \sum p_0}$	先平均（加权）， 后对比 平均指标指数	可做 因素 分析
物值 （总量） 指数	简单指数	综合指数		$V = \frac{\sum q_1 p_1}{\sum q_0 p_0}$	先综合（不加权）， 后对比 物值总量指数	可做 因素 分析
		平均指数		$\frac{1}{n} \sum \frac{q_1 p_1}{q_0 p_0}$	先对比， 后平均（不加权） 很少使用	

4.3 指数的性质

指数的性质介绍：定基指数的特性、指数的测验、基期变动等。

4.3.1 定基指数的特性

定基指数的特性如下。

- 1) 简单综合指数，因为物品的计量或计价单位不同（每千克、每吨、每磅等），其差别可能影响很大。其他指数不受计价单位的影响。
- 2) 加权综合指数中，帕氏指数的权数是变动的（报告期的量）。拉氏指数与罗氏指数的权数是固定的（基期或某期的量），所以罗氏指数又称为“变形拉氏指数”。
- 3) 使用不同的计算（平均）公式，而造成指数的不同称作“型偏”（type bias）。
- 4) 使用不同的权数，而造成指数的不同称作“权偏”（weight bias）。
- 5) 简单平均指数，算术平均指数 A 大于几何平均指数 G ；几何平均指数 G 大于调和平均指数 H ： $A > G > H$ 。

简单算术平均指数 A 的“型偏”是偏大；简单调和平均指数 H 的“型偏”是偏小。

6) 加权综合指数，拉氏指数 P_L, Q_L 通常大于费氏指数 P_F, Q_F ；费氏指数 P_F, Q_F 通常大于帕氏指数 P_P, Q_P 。因为拉氏指数的分子是 $\sum p_1 q_0$ ，即新（报告期）价格乘以旧（基期）数量。但是价格通常是上涨的，根据经济学的理论，价格上涨，则数量降低。所以拉氏权数 q_0 是高估，于是拉氏指数偏高。同理，帕氏指数的权数 q_1 又是偏低，所以帕氏指数通常是较小。指数的比较如表 4-2 和图 4-1 所示。

表 4-5 4.2 节的指数之比较

	物价指数	物量指数
拉氏指数	1.31	0.94
罗氏指数	1.30	0.93
杨格指数	1.30	0.95
费氏指数	1.30	0.93
帕氏指数	1.29	0.92

7) 拉氏指数， $L = A_I = H_{III}$ 。 A_I 是以 $p_0 q_0$ 为权数的加权算数平均指数， H_{III} 是以 $p_0 q_0$ 为权数的加权调和平均指数。

帕氏指数, $P = A_{II} = H_{IV}$ 。 A_{II} 是以 p_0q_0 为权数的加权算数平均指数, H_{IV} 是以 p_1q_1 为权数的加权调和平均指数。

8) 加权平均指数, 在相同的值权之下: 算术指数大于几何指数; 几何指数大于调和指数。

4.3.2 指数的测验

定义 I_{01} 为时间 0 为基期, 时间 1 为报告期的指数。

指数的测验性质如下。

1) 时间互换测验 (time-reversal test): 如果相同公式的指数, 其基期与报告期互换, 而相乘等于 1, 则该指数满足时间互换测验, 即

$$I_{01} \times I_{10} = 1$$

式中: I_{01} ——以 0 期为基期, 报告期为 1 期的指数;

I_{10} ——以 1 期为基期, 报告期为 0 期的指数。

例如: 以 0 期为基期, 1 期为报告期的拉氏物价指数 $P_{L,01} = \frac{\sum p_1 q_0}{\sum p_0 q_0}$

以 1 期为基期, 0 期为报告期的拉氏物价指数 $P_{L,10} = \frac{\sum p_0 q_1}{\sum p_1 q_1}$

因为 $P_{L,01} \times P_{L,10} \neq 1$, 所以, 拉氏指数不符合时间互换测验。

以 1 期为基期, 0 期为报告期的帕氏物价指数 $P_{P,10} = \frac{\sum p_0 q_0}{\sum p_1 q_0}$

但是, $P_{L,01} \times P_{P,10} = 1$ 。

2) 因素互换测验 (factor-reversal test): 若相同公式的物价指数与物量指数, 相乘等于物值指数, 则该指数满足因素互换测验, 即

$$P_{01} Q_{01} = V_{01}$$

式中: P_{01} ——以 0 期为基期, 报告期为 1 期的物价指数;

Q_{01} ——以 0 期为基期, 报告期为 1 期的物量指数;

V_{01} ——以 0 期为基期, 报告期为 1 期的物值指数。

例如: 拉氏物价指数 $P_L = \frac{\sum p_1 q_0}{\sum p_0 q_0}$

拉氏物量指数 $Q_L = \frac{\sum q_1 p_0}{\sum q_0 p_0}$

总值指数
$$V = \frac{\sum q_1 P_1}{\sum q_0 P_0}$$

但是 $P_L \times Q_L \neq V$ ，拉氏物量指数 \times 拉氏物价指数 \neq 物值指数。
所以，拉氏指数不符合因素互换测验。

帕氏物价指数
$$P_p = \frac{\sum P_1 q_1}{\sum P_0 q_1}$$

帕氏物量指数
$$Q_p = \frac{\sum q_1 P_1}{\sum q_0 P_1}$$

同理 $P_p \times Q_p \neq V$ ，帕氏物量指数 \times 帕氏物价指数 \neq 物值指数。
所以，帕氏指数也不符合因素互换测验。

不过 $P_F \times Q_F = V$ ，费氏指数符合因素互换测验。

$Q_L \times P_p = V$ ，拉氏物量指数 \times 帕氏物价指数 = 物值指数（因素分析）。

$Q_p \times P_L = V$ ，帕氏物量指数 \times 拉氏物价指数 = 物值指数（因素分析）。

3) 循环测验（circular test）： m 期环比指数，则相乘等于 1，即

$$I_{01} I_{12} \cdots I_{m-1,m} I_{m0} = 1$$

I_{st} 是以 s 期为基期，报告期为 t 期的指数。

4) 如果指数公式满足循环测验，则一定满足时间互换测验。但是满足时间互换测验，不一定满足循环测验。

5) 费氏指数因为满足时间互换测验及因子互换测验，所以称作“理想指数”。

6) 指数的测验公式，只是参考的性质，并非选择标准。例如拉氏指数，虽然 3 个测验都不满足，但是还是常被采用。

介绍指数的测验的目的是可以多了解指数的定义。

总指数的测验结果如表 4-6 所示。

表 4-6 总指数的测验结果

分类 1	分类 2	分类 3	时间互换	因子互换	循环测验
简单指数	综合指数		○	×	○
	平均指数	算术平均	×	×	×
		几何平均	○	×	○
		调和平均	×	×	×

续表

分类 1	分类 2	分类 3	时间互换	因子互换	循环测验
加权指数	综合指数	拉氏指数	×	×	×
		帕氏指数	×	×	×
		马氏指数	○	×	×
		罗氏指数	○	×	○
		费氏指数	○	○	×
	平均指数	算术平均	×	×	×
		调和平均	×	×	×
		几何平均	×	×	×

4.3.3 基期变动

基期变动是将旧基期的指数改为新基期的指数。其作法是将新基期的旧指数，除以所有的旧指数，就可转换为新的指数。例如：旧指数是 $I_{0t} (t = 1, 2, \dots)$ ，以 0 期为基期，现在要改为以 s 期为基期，则新的指数是

$$I_{st} = \frac{I_{0t}}{I_{0s}}, \quad t = 1, 2, \dots$$

如果指数公式符合循环测验，则基期变动后，新的指数仍合乎原指数的计算公式。例如：简单综合指数变动基期后，用上述除法所得的指数，等于用原数据以新基期，利用简单综合指数公式所得的指数。简单几何平均指数与罗氏指数也符合循环测验

$$\frac{P_{L0,0t}}{P_{L0,0s}} = \frac{\frac{\sum p_t q_a}{\sum p_0 q_a}}{\frac{\sum p_s q_a}{\sum p_0 q_a}} = \frac{\sum p_t q_a}{\sum p_s q_a} = P_{L0,st}$$

但是，如果指数公式若不合乎循环测验，则基期变动后，新的指数不合乎原指数的计算公式。例如：拉氏指数 P_L 变动基期后，用上述除法所得的指数，等于罗氏指数 P_{Lo} 。

$$\frac{\text{拉氏指数基期 0, 报告期 } t}{\text{拉氏指数基期 0, 报告期 } s} = \frac{P_{L,0t}}{P_{L,0s}} = \frac{\frac{\sum p_t q_0}{\sum p_0 q_0}}{\frac{\sum p_s q_0}{\sum p_0 q_0}} = \frac{\sum p_t q_0}{\sum p_s q_0} = P_{Lo,st} = \text{罗氏指数基期 } s, \text{ 报告期 } t$$

拉氏指数基期变动后，新的指数是以 s 期为基期，但是以 q_0 为权数的罗氏指数。

例题 4.3 表 4-7 是将基期 1990 年 1 月的指数, 转换为基期 2012 年 1 月的指数。

表 4-7 基期变动计算表

时间	基期 1990 年 1 月 基期指数 100	基期 2012 年 1 月 基期指数 100
2011 年 1 月	379.7	$379.7 \div 394.5 \times 100 = 96.2$
2011 年 7 月	385.8	$385.8 \div 394.5 \times 100 = 97.8$
2012 年 1 月	394.5	$394.5 \div 394.5 \times 100 = 100$
2012 年 7 月	402.0	$402.0 \div 394.5 \times 100 = 101.9$
2013 年 1 月	407.5	$407.5 \div 394.5 \times 100 = 103.3$
2013 年 7 月	420.5	$420.5 \div 394.5 \times 100 = 106.6$

4.4 指数体系与因素分析

指数体系是一些相关的指数, 在经济的结构上, 所形成的数学关系式, 相关指数的加、减、乘、除的等式如下

总量指数 = 物量指数 × 物价指数

销售额指数 = 销售量指数 × 销售价格指数

总产值指数 = 产量指数 × 产品价格指数

总成本指数 = 产量指数 × 单位产品成本指数

因素分析是根据指数体系, 分析各因素对某一经济指标的影响。

4.4.1 总值指数的因素分析

总值(量)指数 = 拉氏物量指数 × 帕氏物价指数 = 帕氏物量指数 × 拉氏物价指数, 即

$$\frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_1 p_0} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times \frac{\sum q_0 p_1}{\sum q_0 p_0}$$

前一节提到, 拉氏指数和帕氏物价指数不符合因素互换测验, 所以:

总量指数 ≠ 拉氏物量指数 × 拉氏物价指数

因素影响差额的关系式

$$\sum q_1 p_1 - \sum q_0 p_0 = (\sum q_1 p_0 - \sum q_0 p_0) + (\sum q_1 p_1 - \sum q_1 p_0)$$

如果是 3 个以上影响因素 (x, y, z), 例如:

销售利润额 = 销售量 × 单位销售价格 × 销售利润率

原材料支出总额 = 产量 × 单耗 × 单位原材料价格

因素分析如下

$$\frac{\sum x_1 y_1 z_1}{\sum x_0 y_0 z_0} = \frac{\sum x_1 y_0 z_0}{\sum x_0 y_0 z_0} \times \frac{\sum x_1 y_1 z_0}{\sum x_1 y_0 z_0} \times \frac{\sum x_1 y_1 z_1}{\sum x_1 y_1 z_0}$$
$$\sum x_1 y_1 z_1 - \sum x_0 y_0 z_0 = (\sum x_1 y_0 z_0 - \sum x_0 y_0 z_0) + (\sum x_1 y_1 z_0 - \sum x_1 y_0 z_0) + (\sum x_1 y_1 z_1 - \sum x_1 y_1 z_0)$$

例题 4.4 总值指数因素分析数据同表 4-2。

商品名称	数量		单价		权数			
	基期 q_0	报告期 q_1	基期 p_0	报告期 p_1	$q_0 p_0$	$q_1 p_1$	$q_1 p_0$	$q_0 p_1$
甲	50	40	\$ 22	\$ 30	1100	1200	880	1500
乙	2	3	\$ 20	\$ 20	40	60	60	40
丙	80	100	\$ 5	\$ 6	400	600	500	480
Σ	132	143	\$ 47	\$ 56	1540	1860	1440	2020

解答：总值指数 $V = \frac{\sum q_1 p_1}{\sum q_0 p_0} = \frac{40 \times 30 + 3 \times 20 + 100 \times 6}{50 \times 22 + 2 \times 20 + 80 \times 5} = \frac{1860}{1540} = 1.2078$

总值增长绝对数 = 1860 - 1540 = 320

拉氏物量指数 $Q_L = \frac{\sum p_0 q_1}{\sum p_0 q_0} = \frac{1440}{1540} = 0.935$

数量增长绝对数 = 1440 - 1540 = -100

帕氏物价指数 $P_P = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{1860}{1440} = 1.2917$

物价增长绝对数 = 1860 - 1440 = 420

总值（量）指数 = 拉氏物量指数（0.935）× 帕氏物价指数（1.2917）= 1.2078

总值增长绝对数 = 数量增长绝对数（-100）+ 物价增长绝对数（420）= 320 (a)

拉氏物价指数 $P_L = \frac{\sum p_1 q_0}{\sum p_0 q_0} = \frac{2020}{1540} = 1.3117$

物价增长绝对数 = 2020 - 1540 = 480

$$\text{帕氏物量指数 } Q_p = \frac{\sum p_1 q_1}{\sum p_1 q_0} = \frac{1860}{2020} = 0.9208$$

$$\text{数量增长绝对数} = 1860 - 2020 = -160$$

$$\text{总量指数} = \text{拉氏物价指数} (1.3117) \times \text{帕氏物量指数} (0.9208) = 1.2078$$

$$\text{总值增长绝对数} = \text{数量增长绝对数} (-160) + \text{物价增长绝对数} (480) = 320 \quad (\text{b})$$

如图 4-2 所示。

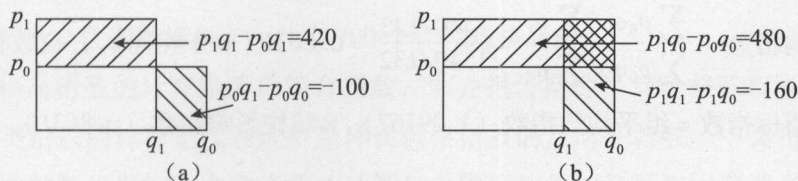


图 4-2 总值增长绝对数的两种计算

4.4.2 平均指数的因素分析

总平均指标指数 = 组平均数指数 × 结构影响指数，即

$$\frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} = \left(\frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_1}{\sum f_1} \right) \times \left(\frac{\sum x_0 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} \right)$$

总平均指标指数 = x_0, x_1 分别以各自权数 f_0, f_1 加权平均，再对比。

组平均数指数 = x_0, x_1 以相同权数 f_1 加权平均，再对比。

结构影响指数 = x_0 以不同权数 f_0, f_1 加权平均，再对比。

$$\frac{\bar{q}_1}{\bar{q}_0} = \frac{\sum p_1 q_1 / \sum p_1}{\sum p_0 q_0 / \sum p_0} = \frac{\sum p_1 q_1 / \sum p_1}{\sum p_1 q_0 / \sum p_1} \times \frac{\sum p_1 q_0 / \sum p_1}{\sum p_0 q_0 / \sum p_0}$$

例题 4.5 同样以表 4-2 的数据，计算物量加权平均指标， x_i 是数量 q_i ，权数 f_i 是价格 p_i 。

$$\text{解答：物量指数以价格作权数，总平均指标指数 } \frac{\bar{q}_1}{\bar{q}_0} = \frac{\sum p_1 q_1 / \sum p_1}{\sum p_0 q_0 / \sum p_0} = \frac{1860/56}{1540/47} = 1.0137$$

$$\text{组平均数指数} = \frac{\sum p_1 q_1 / \sum p_1}{\sum p_1 q_0 / \sum p_1} = \frac{1860/143}{2020/143} = 0.9208 \quad (\text{对于相同权数, } q_i \text{ 变动的影响})$$

$$\text{结构影响指数} = \frac{\sum p_1 q_0 / \sum p_1}{\sum p_0 q_0 / \sum p_0} = \frac{2020/56}{1540/47} = 1.10088 \quad (\text{对于相同 } q_i, \text{ 权数变动的影响})$$

总平均指标指数 = 组平均数指数 (0.9208) × 结构影响指数 (1.10088) = 1.0137

如果以表 4-2 的数据, 计算物价加权平均指标, x_i 是价格 p_i , 权数 f_i 是数量 q_i 。

物价指数以数量作权数, 总平均指标指数 $\frac{\bar{p}_1}{\bar{p}_0} = \frac{\sum p_1 q_1 / \sum q_1}{\sum p_0 q_0 / \sum q_0} = \frac{1860/143}{1540/132} = 1.1149$

组平均数指数 = $\frac{\sum p_1 q_1 / \sum q_1}{\sum p_0 q_1 / \sum q_1} = \frac{1860/143}{1440/143} = 1.29167$ (对于相同权数, q_i 变动的影响)

结构影响指数 = $\frac{\sum p_0 q_1 / \sum q_1}{\sum p_0 q_0 / \sum q_0} = \frac{1440/143}{1540/132} = 0.8631$ (对于相同 q_i , 权数变动的影响)

总平均指标指数 = 组平均数指数 (1.29167) × 结构影响指数 (0.8631) = 1.1149

4.4.3 实质所得指数的应用

指数体系不只是用相乘的, 也可以用相除的。

消费者物价指数可以用来表示“购买力”或“实质所得”(real earnings)。购买力指数或“实质所得指数”等于“平均所得指数”除以“消费者物价指数”, 即

实质所得指数 = 平均所得指数 ÷ 消费者物价指数。

例题 4.6 表 4-8 是将基期 1990 年的平均所得指数, 与基期 1980 年的消费者物价指数, 同时转换为基期 2010 年的指数。然后计算基期为 2010 年的实质所得指数。

表 4-8 物价指数与实质所得

时间/年	平均所得指数 (1990 = 100)	消费者物价 (1980 = 100)	平均所得指数 (2010 = 100)	消费者物价指数 (2010 = 100)	实质所得指数 (2010 = 100)
2011	335.1	152.8	$\frac{335.1}{335.1} \times 100 = 100$	$\frac{152.8}{152.8} \times 100 = 100$	$\frac{100}{100} \times 100 = 100$
2012	351.8	162.8	$\frac{351.8}{335.1} \times 100 = 105.0$	$\frac{162.8}{152.8} \times 100 = 106.5$	$\frac{105.0}{106.5} \times 100 = 0.99$
2013	373.2	176.8	$\frac{373.2}{335.1} \times 100 = 111.4$	$\frac{176.8}{152.8} \times 100 = 115.7$	$\frac{111.4}{115.7} \times 100 = 0.96$
2014	385.9	191.2	$\frac{385.9}{335.1} \times 100 = 115.2$	$\frac{191.2}{152.8} \times 100 = 125.1$	$\frac{115.2}{125.1} \times 100 = 0.92$
2015	394.5	206.9	$\frac{394.5}{335.1} \times 100 = 117.7$	$\frac{206.9}{152.8} \times 100 = 135.4$	$\frac{117.7}{135.4} \times 100 = 0.87$

4.5 指数应用

4.5.1 消费者价格指数

消费者价格指数或称居民消费价格指数 (consumer price index, CPI) 的编制, 将居民消费的商品分为八大类, 每个大类包括若干个中类, 每个中类包括若干个基本分类, 基本分类之下有若干个代表规格品, 大约有 700 个代表规格品。

代表规格品指数的计算是简单综合指数, 即为报告期的平均价格除以基期的平均价格。基本分类指数的计算是代表规格品环比价格指数的几何平均数。中类指数的计算是基本分类指数的加权平均。大类指数的计算是中类指数的加权平均。消费者价格指数的计算是大类指数的加权平均。权重是根据居民家庭用于各种商品和服务的支出额占支出总额的比重。

4.5.2 股价指数

股价指数是用以代表整个股票市场或是某一产业之股价水平的一种指标。股价指数是以指数形式来表示股价水平。由某一时点与基期股价指数的相对大小, 可得知从基期到该时点, 市场上股价变动的幅度。

编算股价指数, 必须在全体上市股票中选取部分代表性的“样本”股票, 使根据样本所计算得之股价指数, 能充分显示其与全体股票的变动具有一致性。

中国台湾证券交易所的“发行量加权股价指数”, 是以样本中各种股票的发行量 (股本), 当作其股价之权数来计算指数, 基本上这个指数是“拉氏链指数”。因为每期 (每天) 可能有新样本上市, 或样本的股本因为新股上市, 而可能有变动。

上证综和指数、美国史坦普尔 500 股票指数 (S&P 500)、纽约证交所综合股票指数和东京证交所综合股票指数的编算方法与发行量加权股价指数一样, 只是采样样本、基期和基期指数有所不同。

4.6 中文统计应用

指数计算和因素分析 (例题 4.4, 例题 4.5)

打开中文统计, 如图 4-3 所示。

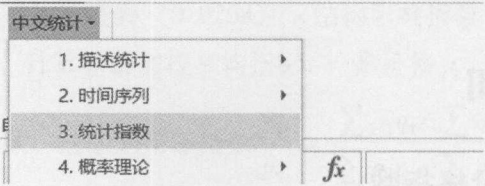


图 4-3 选择“统计指数”

界面如图 4-4 所示。

简单指数(数量或物价指数)										
加权指数(3 个商品)										
加权指数(4 个商品)										
加权指数(5 个商品)										
1	加权指数									
2	商品数目		3							
3										
4			数量		单价		权数			
5	商品名称	计量单位	基期 q_0	报告期 q_1	基期 p_0	报告期 p_1	基期 q_0p_0	报告期 q_1p_1	假定 q_1p_0	假定 q_0p_1
6	甲		50	40	22	30	1100	1200	880	1500
7	乙		2	3	20	20	40	60	60	40
8	丙		80	100	5	6	400	600	500	480
9		Σ	132	143	47	56	1540	1860	1440	2020
10										
11	综合指数						平均指数			
12										
13	物量指数	拉氏	帕氏	罗氏	马氏	费氏	算术	杨格	调和 H	
14	数量指数	0.935065	0.920792	0.931293	0.926966	0.927901	0.935065	0.9517148	0.894482	
15										
16										
17										
18	综合指数						平均指数			
19	物价指数	拉氏	帕氏	罗氏	马氏	费氏	算术	杨格		
20	质量指数	1.311688	1.291667	1.304054	1.302013	1.301639	1.311688	1.3057105	数量	单价
21									个体指数	个体指数
22	总量指数	1.207792							0.8	1.363636
23	总量	=	数量	\times	质量				1.5	1
24	因素分析	总量指数	=	拉氏数量指数	\times	帕氏质量指数			1.25	1.2
25		1.207792	=	0.935065	\times	1.291667			3.55	3.563636
26		总量指数	=	拉氏质量指数	\times	帕氏数量指数				
27		1.207792	=	1.311688	\times	0.920792				
28										

平均指标物量指数				
总平均水平指数	=	组水平变动指数	\times	结构变动指数
$q_1/q_0 = 1.013683$	=	0.920792	\times	1.100881
可变构成指数	=	固定构成指数	\times	结构影响指数
平均指标物价指数				
总平均水平指数	=	组水平变动指数	\times	结构变动指数
$p_1/p_0 = 1.114885$	=	1.291667	\times	0.863137
可变构成指数	=	固定构成指数	\times	结构影响指数
总量差额				
320	=	数量影响	+	质量影响
		-100	+	420

图 4-4 加权指数选择商品数目，在绿色单元格输入数据

4.7 本章流程图

本章流程图如图 4-5 所示。

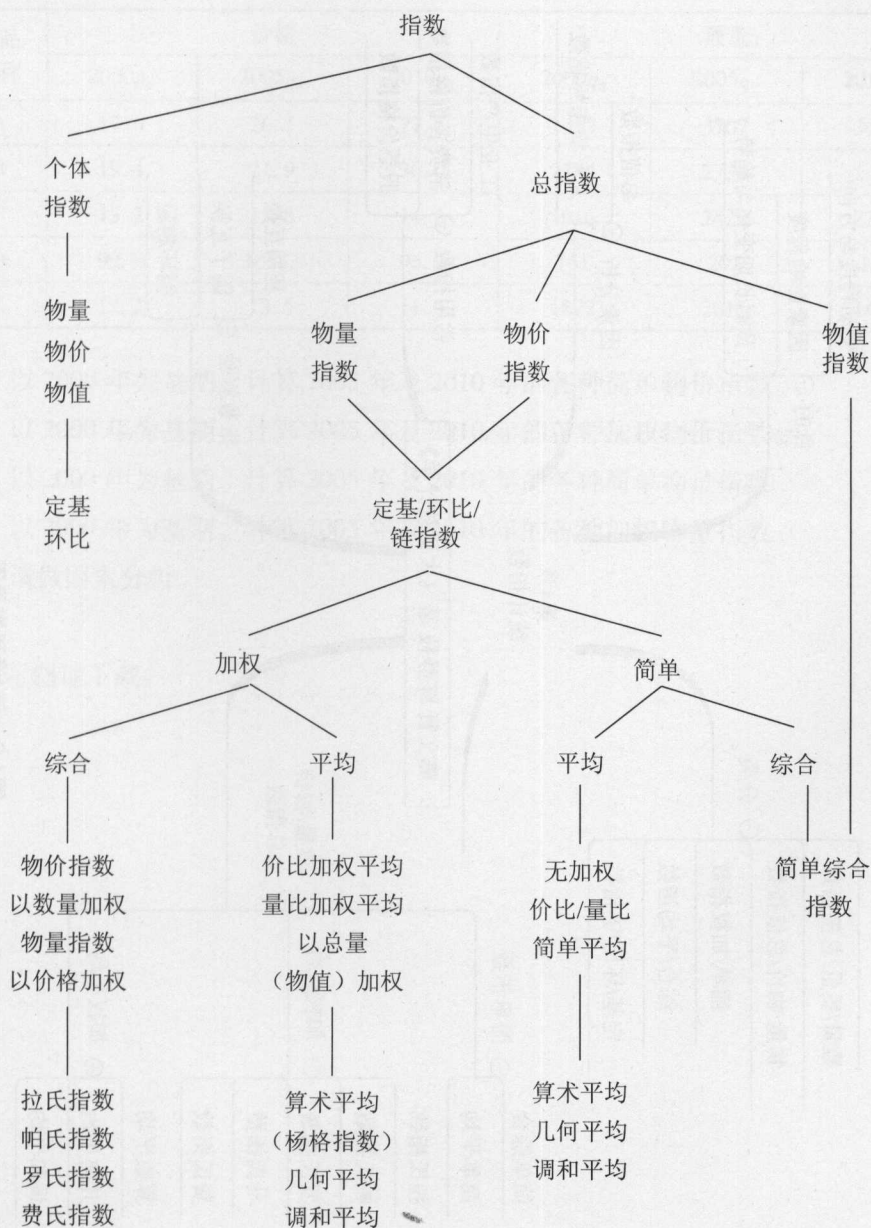


图 4-5 第 4 章的流程图

4.8 本章思维导图

本章思维导图如图 4-6 所示。

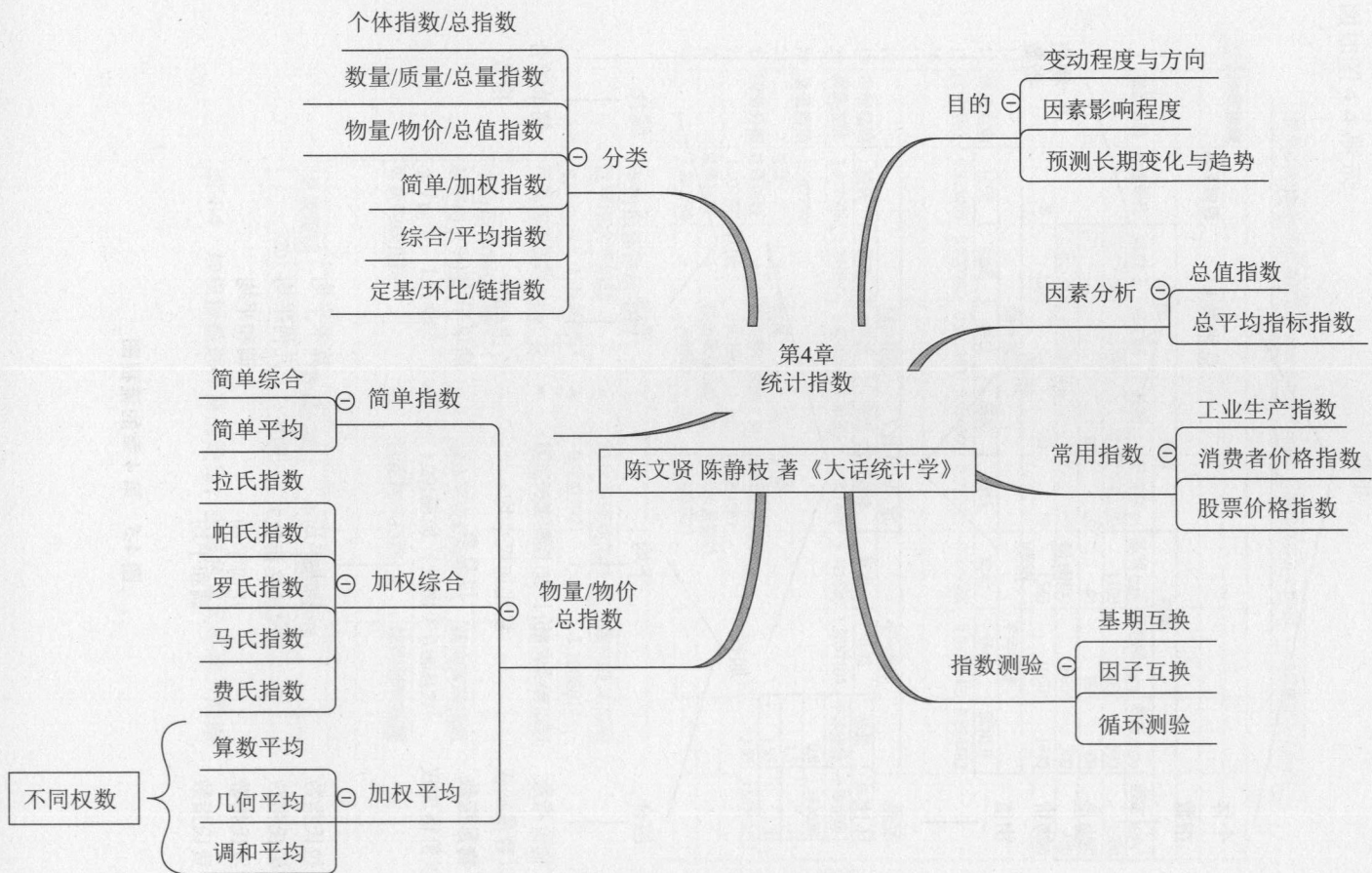


图4-6 第4章思维导图

习题

下列5种商品的价格与交易数量：

商品 名称	价格			数量		
	2000 p_0	2005 p_1	2010 p_2	2000 q_0	2005 q_1	2010 q_2
A	17.0	26.1	27.5	1357	3707	3698
B	19.4	21.9	30.0	2144	2734	2478
C	15.2	15.8	14.5	1916	2420	2276
D	99.3	101.3	96.2	161	202	186
E	12.2	13.5	11.4	1872	2018	1424

- (1) 以2000年为基期，计算2005年及2010年的各种简单物价指数。
- (2) 以2000年为基期，计算2005年及2010年的各种加权物价指数。
- (3) 以2000年为基期，计算2005年及2010年的各种简单物量指数。
- (4) 以2000年为基期，计算2005年及2010年的各种加权物量指数。
- (5) 请做因素分析。

其他习题请下载。



第5章

概率理论

孔明叹曰：“谋事在人，成事在天。不可强也！”

——罗贯中《三国演义》

上帝不掷骰子 (God does not play dice)!

——爱因斯坦 (Albert Einstein)

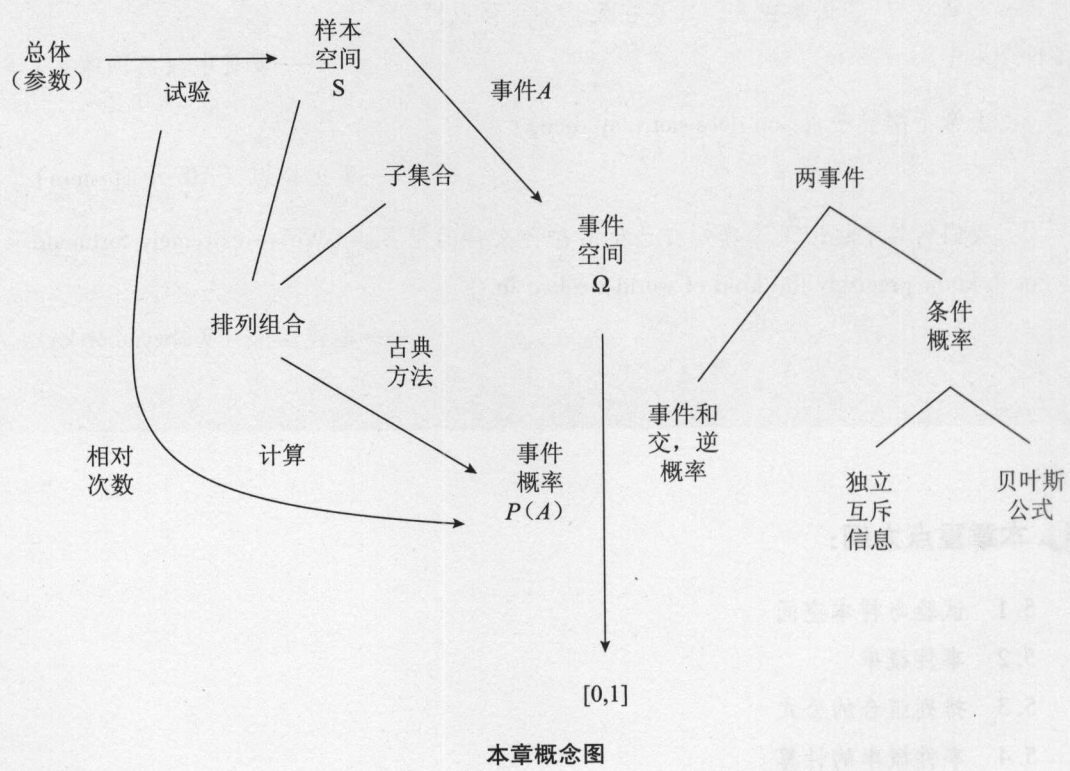
我们何其幸运，无法确知自己生活在什么样的世界。(We're extremely fortunate not to know precisely the kind of world we live in.)

——辛波丝卡 (W. Szymborska)



本章重点大纲:

- 5.1 试验与样本空间
- 5.2 事件概率
- 5.3 排列组合的公式
- 5.4 事件概率的计算
- 5.5 条件概率
- 5.6 独立事件与互斥事件
- 5.7 贝叶斯公式
- 5.8 中文统计应用
- 5.9 本章流程图
- 5.10 本章思维导图



5.1 试验与样本空间

我们面对环境（环境是我们无法掌控）的变化，有6种现象。

1) 确定 (certain) 现象：变量只有一个结果，而且确定会发生。不应该称为“变”量。表达式：一个变量 $x = a$ （方差为0）；两变量（完全相关），圆面积公式 $S = \pi r^2$ 。

2) 循环 (cyclic) 现象：有多种结果，每隔一段时间，有一种结果会确定发生。

3) 随机 (stochastic) 现象：有多种可能的结果，但只有一种结果会发生，不过知道各种结果发生的概率。

4) 模糊 (fuzzy) 现象：有多种可能的结果，只有一种结果会发生，不知道各种结果发生的概率；只知道模糊可能。

5) 不确定 (uncertain) 现象：有多种结果，只有一种结果会发生，不知道各种结果的可能。

6) 混乱或混沌 (chaos) 现象：不知道有那些结果，当然不知道各种结果的可能性。

以学生考试为例，考选择题（单选题，考题是学生无法掌控的），如果一个好学生确定知道答案，这是确定现象。如果学生知道老师的答案顺序（老师为了改题方便），这是循环现象。如果学生不知道确定答案，先删去不可能的答案，再选择概率最大或最有可能的答案，这是随机现象或模糊现象。至于不确定现象，有个笑话：一个学生考试，选择题完全不知道答案，他在铅笔上刻上“1”“2”“3”“4”，因为不知道概率或可能性，所以假定概率相等，于是滚动铅笔回答选择题，很快写完答案，他就睡上一觉，醒来考试还没完，他再滚动起铅笔，老师问他：你在做什么？他回答：我在检查答案。如果老师考问答题，学生完全不会，这是混乱现象。

爱因斯坦说“上帝不掷骰子”，并不是说上帝不赌博或不准赌博，而是说上帝没有概率或不确定性。因为他想要什么就有什么结果，所以其结果是确定的。而且上帝可以控制一切因素，例如掷一个骰子，如果手掌握的位置、力道、方向、距离、高度、角度，甚至温度、气压、情绪等所有的因素都相同，那么掷出骰子的结果会相同（确定结果）。但是，自然界或者人类社会，有许多我们不知道或无法控制的因素，造成了概率或不确定的结果。例如天气台风的预测，或者新产品需求量的预测。

定义 试验 (trial) 是一个过程，其结果是可能的几种情况之一，但是在试验前，不能事先预知结果。

例如下列试验：掷两个骰子；52张扑克牌，抽出5张牌；100件产品（其中有10件

不良品), 抽出 5 件; 抽出一个学生的成绩; 调查一个顾客等候结账的时间。

定义 一个试验结果的每种可能情况, 称作一个基本结果 (elementary outcome), 或称作样本点 (sample point)。

100 件产品 (其中有 10 件不良品), 抽出 5 件的试验, 样本点是 100 件取出 5 件的组合; 抽出一个班级学生的统计学成绩的试验, 样本点是 0 分到 100 分。

例题 5.1 掷一个硬币 (正反两面), 直到出现第一个正面, 则停止。我们用 H 表示正面 (人头 head), T 表示反面 (tail)。样本点: $\{H\}$ 、 $\{TH\}$ 、 $\{TTH\}$ 、 $\{TTTH\}$ 、 $\{TTTTH\}$...

定义 一个试验的“所有”样本点的集合, 称为该试验的样本空间 (sample space), 记作 S 。

定义 一个试验的“部分”样本点的集合, 也就是样本空间的部分集合, 则称为该试验的事件 (event)。

事件是试验结果的一个叙述。例如: 掷两个骰子 (试验), 出现两个都是相同点数 (叙述), 为一个事件 $A: A = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$ 。

定义 一个样本空间“所有”事件的集合, 称为事件空间 (event space), 记作 Ω 。

Ω 的符号在有些统计书中, 定义为样本空间。

定义 不包含任何样本点的“空集合”, 则称为该试验的空事件 (null event), 记作 φ 。

S = 一定会发生 (必然) 的事件, φ = 一定不会发生 (不可能) 的事件。

样本空间 $S = \{\text{所有样本点}\}$, 事件空间 $\Omega = \{S, \varphi, \text{所有的事件}\}$ 。

如果样本空间 S 有 n 个样本点, 则事件空间 Ω 可能有 2^n 个事件。

定义 只包含“一个”样本点的集合, 则称为该试验的基本事件或简单事件。

事件 A 与 B 的和 (并), 记作 $A \cup B$ 或 $A + B$, 本书采用前者符号, 如图 5-1 (a) 所示。

事件 A 与 B 的交 (积), 记作 $A \cap B$ 或 AB , 本书采用前者符号, 如图 5-1 (b) 所示。

事件 A 的逆事件或对立事件, 记作 \bar{A} , 如图 5-1 (c) 所示。

事件 A 与 B 的差, 记作 $A - B = A \cap \bar{B}$ 。

有了集合的定义, 则事件之间的关系, 可以用集合论的文氏图 (venn diagram) 来表示。

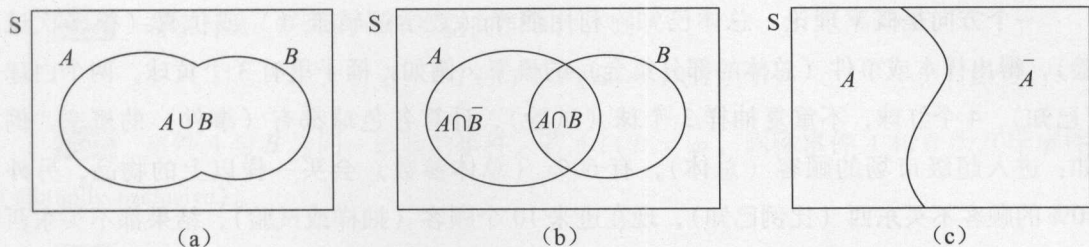


图 5-1 事件的和、交与逆事件

(a) 事件的和; (b) 事件的交; (c) 逆事件

定理 事件的交、和、差的交换律、结合律、分配律。

交换律 $A \cap B = B \cap A$, $A \cup B = B \cup A$

结合律 $(A \cap B) \cap C = A \cap (B \cap C) = A \cap B \cap C$

$(A \cup B) \cup C = A \cup (B \cup C) = A \cup B \cup C$

分配律 $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, $A \cap (\bigcup_{i=1}^n B_i) = \bigcup_{i=1}^n (A \cap B_i)$

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$, $A \cup (\bigcap_{i=1}^n B_i) = \bigcap_{i=1}^n (A \cup B_i)$

De Morgan 律 $\overline{(A \cup B)} = \bar{A} \cap \bar{B}$, $\overline{(\bigcup_{i=1}^n A_i)} = \bigcap_{i=1}^n \bar{A}_i$

$\overline{(A \cap B)} = \bar{A} \cup \bar{B}$, $\overline{(\bigcap_{i=1}^n A_i)} = \bigcup_{i=1}^n \bar{A}_i$

例题 5.2 男女出生概率。(解答见网络资源)

5.2 事件概率

探讨总体与样本之间的性质，分成两个方向，如图 5-2 所示。

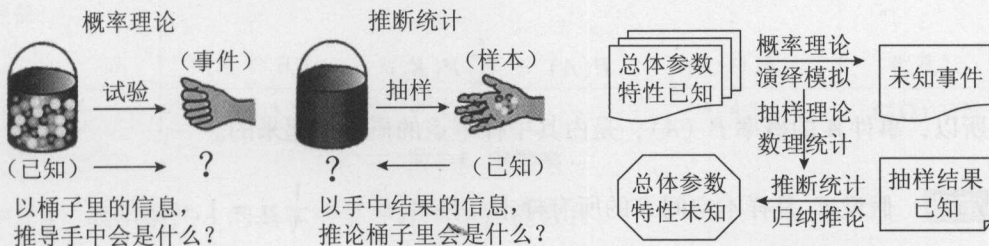


图 5-2 概率和推断统计的不同 (参考: MIT Open Courses)

一个方向是概率理论：总体已知，利用演绎（数学逻辑推导）或仿真（很多次试验），得出样本或事件（总体的部分集合）的概率。例如：桶子里有 3 个黄球，两个白球（已知），4 个红球，不重复抽样 3 个球（试验），计算各色球都有（事件）的概率。例如：进入超级市场的顾客（总体），有 60%（总体参数）会买一样以上的物品，另外 40% 的顾客不买东西（比例已知），现在进来 10 个顾客（抽样或试验），结果都不买东西（事件）的概率是多少。

另一个方向是推断统计：样本已知，推断总体的参数。

定义 如果试验结果，出现某一事件所包含的样本点，则称此事件发生。 $P(A)$ 为事件 A 发生的概率，简称事件 A 的概率（probability of event A ）。

$$P: \Omega \rightarrow [0, 1]$$

要计算事件 A 的概率 $P(A)$ ，利用下列 3 个公理。

公理 1 任何事件 $A: 0 \leq P(A) \leq 1$

公理 2 $P(S) = 1$

公理 3 若事件 A 和 B 是互斥， $A \cap B = \varnothing$ ，则 $P(A \cup B) = P(A) + P(B)$

根据上述公理，可以证明下列定理。（证明省略）

定理 $P(\varnothing) = 0$

定理 $P(\bar{A}) = 1 - P(A)$ ， $P(A) = 1 - P(\bar{A})$

$P(\text{至少出现 1 次}) = P(\text{出现 1 次}) + P(\text{出现 2 次}) + P(\text{出现 3 次}) + P(\text{出现 4 次}) + \dots$

利用逆事件的概率计算： $P(\text{至少出现 1 次}) = 1 - P(\text{都没有出现})$

定理 假设样本空间 S 有 n 个事件 E_1, E_2, \dots, E_n 。 E_i 和 E_j 是互斥： $E_i \cap E_j = \varnothing, \forall i \neq j$ 。

若 $A = \bigcap_{i=1}^n E_i$ ，则

$$P(A) = \sum_{i=1}^n P(E_i)$$

所以，事件 A 的概率 $P(A)$ ，是由其中样本点的概率加起来的。

定理 假设 E_i 是样本空间 S 的所有样本点且 $P(E_i) = \frac{1}{N}, i = 1, \dots, N$ 。

若 $A = \bigcap_{i=1}^n E_i$ ，则 $P(A) = \frac{n}{N}$ 。

定理 若事件 A 、 B 为同一样本空间的事件，则

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

定义 事件 A 与 B 为同一试验的事件，若 $A \cap B = \varnothing$ ，则称事件 A 和 B 为互斥事件 (mutually exclusive)。

定义 样本空间 S 有 n 个事件 A_1, A_2, \dots, A_n 。若 A_i 和 A_j 是互斥, $A_i \cap A_j = \varnothing, \forall i \neq j$ 且 $\bigcap_{i=1}^n A_i = S$ ，则称 $\{A_1, A_2, \dots, A_n\}$ 为 S 的一个完备事件组 (exhaustive events) 或分割 (partition)，如图 5-3 所示。

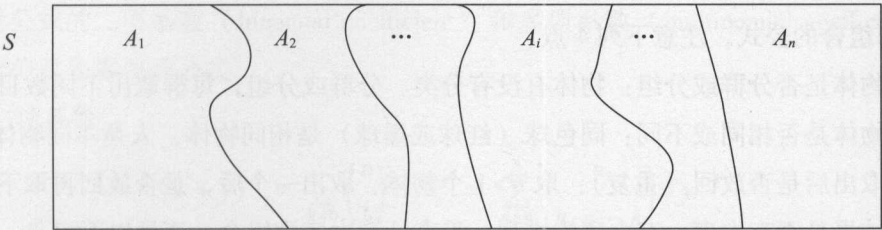


图 5-3 一个完备事件组或分割

$\{A, \bar{A}\}$ 是一个完备事件组或分割。

有的样本空间，可以按照两种不同性质来分割（每一类分别有各种事件），例如：样本空间是一班学生，可以利用其体重与血型两种性质分类（体重有小于 50kg 等事件，血型有 A 型等事件）。将这两种性质，分别为表格的行和列，称作列联表 (contingency table)，如图 5-4 所示。列联表中的每一格表格数据，则填入其样本点的数目。

		性质 1 (分割 1) 完备事件组				
		A_1	A_2	...	A_r	
性质 2 (分割 2) 完备 事件组	B_1	$\#(A_1 \cap B_1)$	$\#(A_2 \cap B_1)$...	$\#(A_r \cap B_1)$	$\#(B_1)$
	B_2	$\#(A_1 \cap B_2)$	$\#(A_2 \cap B_2)$...	$\#(A_r \cap B_2)$	$\#(B_2)$
	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
	B_s	$\#(A_1 \cap B_s)$	$\#(A_2 \cap B_s)$...	$\#(A_r \cap B_s)$	$\#(B_s)$
		$\#(A_1)$	$\#(A_2)$...	$\#(A_r)$	$\#(\Omega) = N$

图 5-4 列联表

$\#(A)$ 表示事件 A 的基本结果 (样本点) 的数目。

从列联表中，计算事件的概率。例如

$$P(A_1) = \frac{\#(A_1)}{N}$$

$$P(A_1 \cap B_1) = \frac{\#(A_1 \cap B_1)}{N}$$

5.3 排列组合的公式

排列组合的公式，可以计算样本空间的样本点的数目，以及事件中的样本点的数目。然后两者相除，就是事件的概率。

排列组合的公式，注意下列 4 点。

- 1) 物体是否分群或分组：物体有没有分类、分群或分组，每群取出不同数目。
- 2) 物体是否相同或不同：同色球（红球或黑球）是相同物体，人是不同物体。
- 3) 取出后是否放回（重复）：取 $r > 1$ 个物体，取出一个后，是否放回再取下一个。
- 4) 结果是否有次序：有次序为排列，没有计算次序为组合，请见以下定义。

定义 若有 r 个不同物体，其先后次序不同，而区别为不同之样本点者，称为排列。

定义 若有 r 个不同物体，不管其先后次序，形成一个集合，而为一个样本点，称为组合。

定理（乘法公式）若有 m 群物体，第 1 群有 k_1 个“不同”物体，第 2 群有 k_2 个“不同”物体，……，第 m 群有 k_m 个“不同”物体。如果一个试验，是从 m 个群中，每一群取出一件物体，则这个试验的“组合”样本点的数目是 $k_1 \times k_2 \times \cdots \times k_m$ 。

乘法公式称为“基本计数原则”。

定理（排列公式）若一个试验，是从一群 n 个“不同”物体，取出 r 个“排列”，取出后“不放回”，则这个试验的“排列”样本点的数目，记作 P_r^n ，其公式为

$$P_r^n = \frac{n!}{(n-r)!}$$

定理（组合公式）若一个试验，是从一群 n 个“不同”物体，取出 r 个“组合”，取出后“不放回”，则这个试验的“组合”样本点的数目，记作 C_r^n 或 $\binom{n}{r}$ ，其公式为

$$C_r^n = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

定理 (多项公式, 相同物体排列公式) 若一个试验, 是 n 个物体排列, 其中有 n_1 是“相同”的第一类物体, 有 n_2 是“相同”的第二类物体, 依此类推, 有 n_k 是“相同”的第 k 类物体, $n_1 + n_2 + \cdots + n_k = n$, 则这个试验的“排列”样本点的数目是

$$\binom{n}{n_1, n_2, \dots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}$$

组合公式是排列 (不同物体) 变组合; 相同物体排列 (多项) 公式是组合 (相同物体) 变排列。其步骤互逆, 两个公式结果相同, 即

$$C_r^n = \binom{n}{r} = \frac{n!}{r!(n-r)!} = \binom{n}{r, n-r} = \binom{n}{n-r} = C_{n-r}^n$$

数学公式的二项系数 (binomial coefficient) 和多项系数 (multinomial coefficient), 是利用组合公式和多项列公式。

1. 二项系数

$$(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k} = \sum_{k=0}^n \binom{n}{k, n-k} x^k y^{n-k}$$

2. 多项系数

$$(x_1 + x_2 + \cdots + x_k)^n = \sum_{n_1+n_2+\cdots+n_k=n} \binom{n}{n_1, n_2, \dots, n_m} x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}$$

定理 (超几何公式) 若物体分成 k 类, 第 1 类有 n_1 个“不同”物体, 第 2 类有 n_2 个“不同”物体, ……第 k 类有 n_k 个“不同”物体。一个试验, 是从第 1 类 n_1 个“不同”物体, 取出 r_1 个, 从第 2 类 n_2 个“不同”物体, 取出 r_2 个, ……从第 k 类 n_k 个“不同”物体, 取出 r_k 个, 取出后“都不放回”, 取出的物体作“组合”。则这个试验的样本点的数目是

$$C_{r_1}^{n_1} C_{r_2}^{n_2} \cdots C_{r_k}^{n_k} = \binom{n_1}{r_1} \binom{n_2}{r_2} \cdots \binom{n_k}{r_k} = \frac{n_1! n_2! \cdots n_k!}{r_1! r_2! \cdots r_k! (n_1 - r_1)! (n_2 - r_2)! \cdots (n_k - r_k)!}$$

超几何公式是利用乘法律与组合公式。

排列和组合的公式如表 5-1 所示。

表 5-1 排列和组合的公式

	只有一类 (组)		有多类 (组)	
	n 个不同物体		每类不同物体	每类相同物体
	取出 r 个不放回	取出 r 个放回 (重复)	不放回	不放回
排列	P_r^n	n^r 重复排列	\times	多项公式
组合	C_r^n	C_r^{n+r-1} 重复组合	超几何公式	\times

组合公式的性质如下。

1. $C_r^n \times r! = P_r^n$, $C_r^n = \frac{P_r^n}{r!}$
2. $C_r^n = C_{n-r}^n$ (左右对称)
3. $C_0^n + C_1^n + C_2^n + \cdots + C_n^n = 2^n$
4. $C_r^n + C_{r+1}^n = C_{r+1}^{n+1}$

帕斯卡尔三角形如图 5-5 所示。

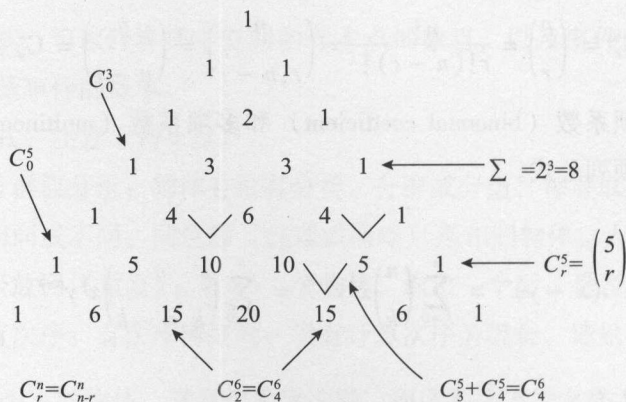


图 5-5 帕斯卡尔三角形

例题 5.3 乘法公式

一套音响要三个部分：收音机，CD 雷射唱盘，喇叭；录音机可有可无。现在有 10 种收音机，8 种 CD，5 种喇叭，3 种录音机。请问：音响的组合结果有几种选择？

解答：音响的组合结果的选择有

$$10 \times 8 \times 5 \times (3 + 1) = 1600$$

例题 5.4 排列公式

一个学会有 30 个会员，现在要选出 1 个会长，1 个副会长，1 个秘书长，1 个财务长。请问：可能的选举结果有几种？

解答：可能的选举结果有

$$P_4^{30} = \frac{30!}{26!} = 30 \times 29 \times 28 \times 27 = 657720$$

例题 5.5 组合公式

一个学会有 30 个会员，现在要选出 6 个会员当作委员会委员。请问：委员会的组成方式有几种？如果有主任委员，则可能的选举结果有几种？

解答：委员会的组成方式有

$$C_6^{30} = \binom{30}{6} = \frac{30!}{6!24!} = \frac{30 \times 29 \times 28 \times 27 \times 26 \times 25}{6 \times 5 \times 4 \times 3 \times 2 \times 1} = 593775$$

如果有主任委员，则可能的选举结果有

$$593775 \times 6 = 3562650$$

例题 5.6 多项公式

STATISTIC 的 9 个字母排列成的字（不管有没有意义）有多少个？

解答：

$$\binom{9}{3,2,2,1,1} = \frac{9!}{3!2!2!1!1!} = \frac{9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{1 \times 2 \times 3 \times 2 \times 2 \times 1 \times 1} = 9 \times 8 \times 7 \times 6 \times 5 = 15120$$

例题 5.7 多项公式

一班有 12 个学生，教授决定要给 2 个“A”，2 个“B”，5 个“C”，3 个“D”。请问：12 个人可能的分数有几种？

解答：

$$\binom{12}{2,2,5,3} = \frac{12!}{2!2!5!3!} = \frac{12 \times 11 \times 10 \times 9 \times 8 \times 7 \times 6}{2 \times 2 \times 3 \times 2 \times 1} = 12 \times 11 \times 10 \times 9 \times 2 \times 7 = 166320$$

例题 5.8 超几何公式

一个班级的篮球选手有 10 个学生，其中有 6 个男生，4 个女生。现在要选出 5 个选手，一定要有 3 男 2 女。请问：球队的组成方式有几种？

解答：球队的组成方式有

$$C_3^6 \times C_2^4 = \binom{6}{3} \times \binom{4}{2} = \frac{6!}{3!(6-3)!} \times \frac{4!}{2!(4-2)!} = 120$$

例题 5.9 重复组合

有 4 种水果：苹果、橘子、梨子、香蕉，可重复取出 12 个，请问有多少可能的选择？

解答：

$$n=4, r=12, n+r-1=15$$

$$\binom{n+r-1}{r} = \binom{15}{12} = \binom{15}{3} = \frac{15 \times 14 \times 13}{1 \times 2 \times 3} = 455$$

1 到 15 个取出 3 个为间隔是一种可能的选择，例如取 3, 11, 12 为间隔，即

1, 2, ③, 4, 5, 6, 7, 8, 9, 10, ⑪, ⑫, 13, 14, 15

表示 2 个苹果、7 个橘子、0 个梨子、3 个香蕉。

例题 5.10 重复排列

一个骰子掷出两次，请问样本点有多少可能？

解答：一个骰子有 6 面，样本点有 $6 \times 6 = 36$ ，这是重复排列。如果用重复组合，则有 $C_2^{6+2-1} = C_2^7 = 21$ ，但是样本点的概率不等， $\{1, 1\}$ 的概率是 $1/36$ ， $\{1, 2\}$ 的概率是 $2/36$ 。

因此，重复组合不适用于下述古典方法的事件概率计算。

5.4 事件概率的计算

事件 A 的概率 $P(A)$ 的计算，如果样本空间很简单，可以列出其样本点，还有下列 3 种方式：古典方法或逻辑推导、相对次数或模拟仿真、主观判断。前两者（逻辑推导和相对次数）计算出的概率，称作客观概率；主观判断得到的概率，称作主观概率。

5.4.1 古典方法

古典方法是用逻辑推导出来的概率，又称为古典概率或先天概率。通常，逻辑推导要利用到排列组合的公式。假设每个样本点的概率相同，则

$P(A)$ = 事件 A 的样本点的数目 \div 样本空间的样本点的数目

$P(A)$ = 事件 A 的样本点的数目 \times 样本点的概率

以上，事件 A 的样本点的数目以及样本空间的样本点的数目是利用（排列组合）公式计算而不是列出样本点。

如果分母用排列公式（有排列次序），则分子也要用排列公式。如果分母用组合公式（不考虑次序），则分子也要用组合公式。

例题 5.11 一个袋子中有 10 个球：5 个红球，3 个白球，2 个绿球。取出 4 个球，每次取出一球后不放回。请问 4 球中，各色球都有（ A ）之概率 $P(A)$ 。

解答：样本空间的样本点的数目 = C_4^{10} ，利用超几何公式

$$P(A) = \frac{\binom{5}{2} \times \binom{3}{1} \times \binom{2}{1}}{\binom{10}{4}} + \frac{\binom{5}{1} \times \binom{3}{2} \times \binom{2}{1}}{\binom{10}{4}} + \frac{\binom{5}{1} \times \binom{3}{1} \times \binom{2}{2}}{\binom{10}{4}} = \frac{1}{2}$$

分母（组合公式）和分子（超几何公式）都把每个色球，当作“不同”物体。

5.4.2 相对次数

假设事件 A 是一个试验的事件，若该试验重复 N 次，而事件 A 出现 n 次，则事件 A 的概率 $P(A) = \frac{n}{N}$ 。

例题 5.12 袋子中：5 个红球，3 个白球，2 个绿球。取出 4 个球不放回，请问各色球都有之概率。

解答：利用模拟，实际用袋子装 10 个色球，或用计算机仿真。100 次试验，出现不同色球的情况如表 5-2 所示，有星号 “*” 者，表示各色球都有的事件。

表 5-2 模拟试验的结果

事件			模拟（仿真）出现频数
红色	白色	绿色	
4	0	0	1
3	1	0	11
3	0	1	6
2	2	0	13
2	1	1	24 *
2	0	2	7
1	3	0	4
1	2	1	20 *
1	1	2	10 *
0	3	1	2
0	2	2	2
模拟（仿真）次数			100

根据 100 次模拟试验的结果，可得

$$P(A) = \frac{24 + 20 + 10}{100} = 0.54$$

大数法则（law of large number）：相当多次数 N 重复试验的结果，相对次数的概率会近似古典方法或逻辑推导的概率。

因为计算机的快速计算，利用计算机模拟或仿真（simulation），使 N 相当大，可以很方便地计算相对次数，如图 5-6 所示。

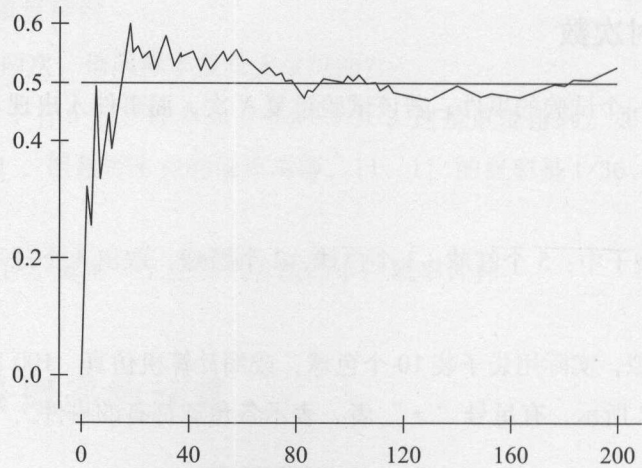


图 5-6 模拟各色球都出现的概率与次数

5.4.3 主观判断

如果试验的样本点太复杂无法用逻辑推导，而且仿真试验有困难，则事件 A 的概率 $P(A)$ 的计算，可能要借助某人或某些人的经验和直觉，利用主观的判断。

例题 5.13 如图 5-7 所示，若正方形 S 的面积是 1，则 A 的面积是多少？请分别用逻辑推导、相对次数、主观判断来估计。

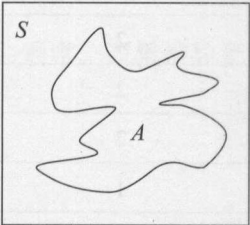


图 5-7 例题 5.15

解答：

- 1) 逻辑推导（微积分）：将 S 分成 $10 \times 10 = 100$ 个正方形格子，计算 A 占有多少格子。用 $20 \times 20 = 400$ 个格子，会更准确。
- 2) 逻辑推导（物理学）：用一块厚纸板，剪下 S 称其质量，再剪下 A 称其质量，比较两者质量。
- 3) 相对次数：放大图形，对 S 随机射 100 个飞镖，假设每镖都平均落在 S ，计算 A 有多少飞镖。
- 4) 主观判断：目测法。

5.5 条件概率

条件概率（conditional probability） $P(A|B)$ ，是在事件 B 确定发生的情况下，事件 A 发生的概率。条件概率是缩小样本空间，以事件 B 为样本空间，计算事件 $A \cap B$ 在 B 的概

率。条件概率主要应用在判断两个事件的独立或相依，或修正事件发生后条件概率的计算。

定义 若事件 A 与 B 为同一试验的事件，且 $P(B) \neq 0$ ，则条件概率 $P(A|B) = \frac{P(A \cap B)}{P(B)}$ ，如图 5-8 所示。

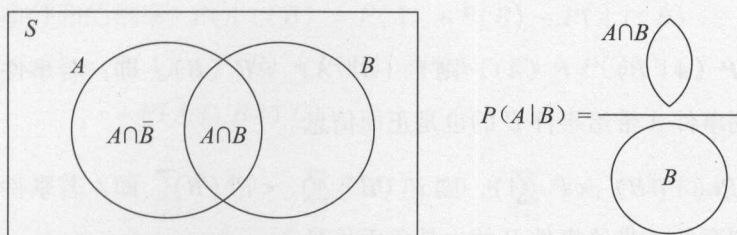


图 5-8 条件概率集合表示图

同理，若 $P(A) \neq 0$ ，则 $P(B|A) = \frac{P(A \cap B)}{P(A)}$ 。

5.5.1 正面信息、负面信息与无信息

有时候对于某些要研究的事件主题 A ，在已经知道部分信息（partial information），例如事件 B 的情况下，研究该主题事件 A 的概率，即为条件概率 $P(A|B)$ 。

若 $P(A|B) > P(A)$ ，或 $P(A \cap B) > P(A)P(B)$ ，则事件 B 带给事件 A 的是正面信息（positive information），于是已知 B 事件，可以增加 A 事件的投资（赌注）。

若 $P(A|B) < P(A)$ ，或 $P(A \cap B) < P(A)P(B)$ ，则事件 B 带给事件 A 的是负面信息（negative information），于是已知 B 事件，应该减少 A 事件的投资（赌注）。

若 $P(A|B) = P(A)$ ，则事件 B 带给事件 A 的是无信息（no information），无信息是无关信息，没有关系的信息，也就是两个事件 A 与 B 是独立的（independent）。已知 B 事件发生，但不影响 A 事件发生的概率。

若事件 A 与 B 为互斥事件 $A \cap B = \varnothing$ ，则 $P(A|B) = 0 < P(A)$ ，事件 B 带给事件 A 的是（完全）负面信息，有了 B 事件就没有 A 事件（不可能发生），同理，有了 A 就没有 B 。已知 B 事件发生，应该停止 A 事件的投资（赌注）。

在数据采勘（data mining）的购物篮分析，可以用条件概率检查，买什么东西的顾客，会再买什么东西。例如买尿片的，会买啤酒， $P(B \text{ 买啤酒} | A \text{ 买尿片}) > P(B \text{ 买啤酒})$ 。在赌场赌 21 点，利用算牌的方法（10 点以上大牌或 2~6 点的小牌），如果剩下未出的牌，大牌很多（事件 B ），庄家爆牌（事件 A ）的概率变大，就下大赌注。

表 1-2 的泰坦尼克号存活数据， S = 船上所有人员， A = 存活人员， B = 头等舱旅客，

C = 二等舱旅客, D = 三等舱旅客, E = 船员组员, 所有人员存活率 $P(A) = 0.32$, 头等舱旅客存活率 $P(A|B) = 0.60$, 二等舱旅客存活率 $P(A|C) = 0.41$, 三等舱旅客存活率 $P(A|D) = 0.24$, 船员组员存活率 $P(A|E) = 0.24$ 。所以, 头等舱旅客、二等舱旅客或女性身份对存活事件是正面信息; 三等舱旅客、船员组员或男性身份对存活事件是负面信息。

定理 若 $P(A|B) > P(A)$, 则 $P(B|A) > P(B)$ 。即: 若事件 B 带给事件 A 是正面信息, 则事件 A 带给事件 B 的也是正面信息。

定理 若 $P(A|B) < P(A)$, 则 $P(B|A) < P(B)$ 。即: 若事件 B 带给事件 A 是负面信息, 则事件 A 带给事件 B 的也是负面信息。

若 A (公司利空消息) 对 B (买进股票) 是负面信息, 则 A 对 B 的逆事件 \bar{B} 是正面信息。
 A, \bar{A}, B, \bar{B} 的关系是图 5-9 中 3 个表格之一, 事件独立的关系请见下一节。每个表格有 4 (2×2) 组事件, 只要一组关系确定, 另外三组关系也可以确定。

(1)	B	\bar{B}	(2)	B	\bar{B}	(3)	B	\bar{B}
A	正	负	A	负	正	A	独立	独立
\bar{A}	负	正	\bar{A}	正	负	\bar{A}	独立	独立
(a)			(b)			(c)		

图 5-9 A, \bar{A}, B, \bar{B} 事件的关系

例题 5.14 (见网络资源)

例题 5.15 (见网络资源)

例题 5.16 两个骰子, 已经知道掷出两个骰子都是黑色 (2, 3, 5, 6) (事件 A), 则点数和为 “12” (事件 B) 之概率 $P(B|A)$ 是多少? 你押注 “点数和 12”; 要不要加码?

解答: 掷出骰子的情况如表 5-3 所示。

表 5-3 掷出骰子的情况

		两个骰子点数和 B_j											
		2	3	4	5	6	7	8	9	10	11	12	$\#(A_j)$
骰子 颜色 A_i	2 红 {1, 4}	1	0	0	2	0	0	1	0	0	0	0	4
	1 红 1 黑	0	2	2	0	4	4	0	2	2	0	0	16
	2 黑 {2, 3, 5, 6}	0	0	1	2	1	2	4	2	1	2	1	16
$\#(B_i)$		1	2	3	4	5	6	5	4	3	2	1	36

$P(B) = \frac{\#(B)}{\#(S)} = \frac{1}{36}$, $P(B|A) = \frac{\#(A \cap B)}{\#(A)} = \frac{1}{16}$, $P(B|A) > P(B)$, 则 A 给 B 是正面信息。

定理 若事件 A 、 B 、 C 为同一样本空间的事件, 则

加法律 (事件和的概率) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum_{i < j} P(A_i \cap A_j \cap A_k) - \sum_{i < j < k < l} P(A_i \cap A_j \cap A_k \cap A_l) + \cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n)$$

乘法律 (事件交的概率) $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B) = P(B)P(C|B)P(A|B \cap C) = P(C)P(A|C)P(B|A \cap C)$

若 $P(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$, 则

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) =$$

$$P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)P(A_4|A_1 \cap A_2 \cap A_3) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})$$

例题 5.17 我们来交易吧 (Let's Make A Deal) (Monty Hall 问题)

1963 年开始, 美国电视游戏节目 (主持人 Monty Hall): 有 3 个门, 其中一个门后面有一辆汽车, 两个门有只羊。你选一个门, 暂时不打开。主持人打开一个有羊的门, 主持人问你: “要不要换门?” 请问: 这时你换或不换? 得到汽车的概率各多少?

解答: 1. 选择 “换门” 策略

如果: L_1 表示第一次选择是错误 (羊), W_1 表示第一次选择是正确 (车); L_2 表示第二次选择 (交换) 是错误, W_2 表示第二次选择 (交换) 是正确, 则各种情况的概率如图 5-10 所示。

L_1	$P(L_1)=2/3$	$\frac{P(L_2)P(L_2 L_1)=0}{P(W_2)P(W_2 L_1)=1}$	$P(L_1 \cap L_2)=0$
			$P(L_1 \cap W_2)=2/3$
W_1	$P(W_1)=1/3$	$\frac{P(L_2)P(L_2 W_1)=1}{P(W_2)P(W_2 W_1)=0}$	$P(W_1 \cap L_2)=1/3$
			$P(W_1 \cap W_2)=0$

图 5-10 选择 “换门” 策略时各种情况的概率

赢汽车的概率 $P(W_2)$ 为

$$\begin{aligned} P(W_2) &= P(L_1 \cap W_2) + P(W_1 \cap W_2) \\ &= P(L_1 \cap W_2) \\ &= P(W_2|L_1)P(L_1) = 1 \times (2/3) = 2/3 \end{aligned}$$

2. 选择“不换”策略

如果： L_1 表示第一次选择是错误（羊）， W_1 表示第一次选择是正确（车）； L_2 表示第二次选择（不换）是错误， W_2 表示第二次选择（不换）是正确，则各种情况的概率如图 5-11 所示。

L_1	$P(L_2)P(L_2 L_1)=1$	$P(L_1 \& L_2)=2/3$
$P(L_1)=2/3$	$P(W_2)P(W_2 L_1)=0$	$P(L_1 \& W_2)=0$
W_1	$P(L_2)P(L_2 W_1)=0$	$P(W_1 \& L_2)=0$
$P(W_1)=1/3$	$P(W_2)P(W_2 W_1)=1$	$P(W_1 \& W_2)=1/3$

图 5-11 选择“不换”策略时各种情况的概率

赢汽车的概率 $P(W_2)$ 为

$$\begin{aligned} P(W_2) &= P(L_1 \cap W_2) + P(W_1 \cap W_2) \\ &= P(W_1 \cap W_2) = P(W_2 | W_1)P(W_1) = 1 \times (1/3) = 1/3 \end{aligned}$$

简单地说，选择换门，原来得到“车子”的（1/3 概率）变成得到“羊”，原来得到“羊”的（2/3 概率）变成得到“车子”。所以，选择换门，得到“车子”的概率是 2/3。

5.6 独立事件与互斥事件

定义 事件 A 与 B 为同一试验的事件，若 $A \cap B = \varnothing$ ，则称事件 A 和 B 为互斥事件 (mutually exclusive)。

互斥事件：两个事件的交集为空集合 \varnothing ，所以 $P(A \cap B) = 0$ 。

定义 若事件 A_1, A_2, \dots, A_n 为 n 个事件，对于任何两个事件： $A_i \cap A_j = \varnothing, i \neq j$ ，则称事件 A_1, A_2, \dots, A_n 为互斥事件。

互斥事件是任何两个以上事件交集是空集合，即两两互斥。

定理 若事件 A 和 B 为互斥事件，则 $P(A \cup B) = P(A) + P(B)$ 。

定义 事件 A 与 B 为同一试验的事件，若下列两条件之一成立：

$$1. P(A|B) = P(A)$$

$$2. P(B|A) = P(B)$$

则称事件 A 和 B 为独立事件 (independent events)。

独立事件：一个事件发生的概率，不影响另一个事件发生的概率。

定理 若且唯若事件 A 和 B 为独立事件, 则 $P(A \cap B) = P(A)P(B)$ 。

定理 若事件 A 和 B 为独立事件, 则 A 和 \bar{B} 为独立事件、 \bar{A} 和 B 为独立事件、 \bar{A} 和 \bar{B} 为独立事件。

图 5-12 中的 4 (2×2) 组事件, 只要一组独立, 另外三组也会是独立。由此推广, ($m \times n$) 组事件的列联表, 如果列联表有 $(m-1) \times (n-1)$ 组事件独立, 则所有组也会是独立。这就是所谓“自由度”。(请见第 14 章)

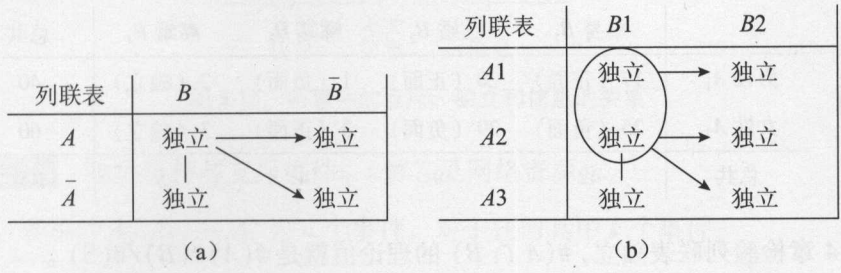


图 5-12 列联表的独立

例题 5.18 甲公司的 100 个员工, 性别与婚姻状况的列联表, 如表 5-4 所示。

A_1 = 男性员工, A_2 = 女性员工
 B_1 = 单身员工, B_2 = 已婚员工, B_3 = 鳏寡员工, B_4 = 离婚员工

表 5-4 性别与婚姻状况的列联表

		性质 2 (婚姻状态)				
		单身 B_1	已婚 B_2	鳏寡 B_3	离婚 B_4	总共
性质 1 (性别)	男性 A_1	17	20	1	2	40
	女性 A_2	25	29	3	3	60
	总共	42	49	4	5	100

- (1) A_1 与 A_2 是否独立事件?
- (2) A_1 与 B_2 是否独立事件?
- (3) A_2 与 B_4 是否独立事件?

解答:

- (1) A_1 与 A_2 是互斥事件, 且 $P(A_1) > 0, P(A_2) > 0$, 所以不是独立事件。
- (2) $P(A_1 | B_2) = \frac{P(A_1 \cap B_2)}{P(B_2)} = \frac{20/100}{49/100} \neq P(A_1) = \frac{40}{100}$, A_1 与 B_2 不是独立事件。
- (3) $P(A_2 | B_4) = \frac{P(A_2 \cap B_4)}{P(B_4)} = \frac{3/100}{5/100} = P(A_2) = \frac{60}{100}$, A_2 与 B_4 是独立事件。

如果 $\#(A)$ = 事件 A 的样本点数目, $\#(S)$ = 样本空间 S 的样本点数目。
若 A 和 B 是列联表的事件, $\#(A \cap B)\#(S) = \#(A)\#(B)$, 则 A 和 B 是独立。
若 A 和 B 是列联表的事件, $\#(A \cap B)\#(S) > \#(A)\#(B)$, 则 A 和 B 是正面信息。
若 A 和 B 是列联表的事件, $\#(A \cap B)\#(S) < \#(A)\#(B)$, 则 A 和 B 是负面信息。
婚姻状态与性别的关系如表 5-5 所示。

表 5-5 婚姻状态与性别的关系

		性质 2 (婚姻状态)				
性 质 1 (性 别)		单身 B_1	已婚 B_2	鳏寡 B_3	离婚 B_4	总共
	男性 A_1	17 (正面)	20 (正面)	1 (负面)	2 (独立)	40
	女性 A_2	25 (负面)	29 (负面)	3 (正面)	3 (独立)	60
	总共	42	49	4	5	100

在第 14 章检验列联表独立, $\#(A \cap B)$ 的理论值就是 $\#(A)\#(B)/\#(S)$ 。

定理 若 $P(B|A) = P(B|\bar{A})$, 则 A 和 B 为独立事件。

证明: $P(B|A) = P(B|\bar{A})$

$$\Rightarrow P(A \cap B)/P(A) = P(\bar{A} \cap B)/P(\bar{A}) = [P(B) - P(A \cap B)]/[1 - P(A)]$$
$$\Rightarrow P(A \cap B)[1 - P(A)] = P(A)[P(B) - P(A \cap B)]$$
$$\Rightarrow P(A \cap B) - P(A)P(A \cap B) = P(A)P(B) - P(A)P(A \cap B)$$
$$\Rightarrow P(A \cap B) = P(A)P(B)$$

事件 A 和 B 只有 3 种情况: ①独立事件; ②互斥事件; ③非独立事件且非互斥事件。
如果 $P(A) > 0, P(B) > 0$, 且 A, B 是独立事件, 则 $P(A \cap B) = P(A)P(B) > 0$, 所以 A, B 不是互斥事件。

如果 $P(A) > 0, P(B) > 0$, 且 A, B 是互斥事件, 则 $P(A|B) = P(A \cap B)/P(B) = 0 \neq P(A)$, 所以 A, B 不是独立事件。

因此, 如果 $P(A) > 0, P(B) > 0, A, B$ 是独立事件, 就不可能是互斥事件。

如果 A, B 是互斥事件, 就不可能是独立事件。

图 5-13 的集合图的关系, A, B 两个事件, 从互斥 $A \cap B = \varnothing$ 开始; 若 $A \cap B$ 不为空集合但很小, 是负面信息; $A \cap B$ 越大, 成为独立 [$A:S = (A \cap B):B$]; 然后 $A \cap B$ 更大, 成为正面信息。最后是 $A \subseteq B$ 或 $B \subseteq A$: 若 $A \subseteq B$, 则 $P(A|B) = P(A)/P(B) > P(A)$; 若 $B \subseteq A$, 则 $P(A|B) = 1$ 。所以, 两事件的交 $A \cap B$ 越大, 越是成为正面信息; 两事件的交 $A \cap B$ 越小, 越是成为负面信息; 两事件互斥, $P(A|B) = 0$, 当然是负面信息。

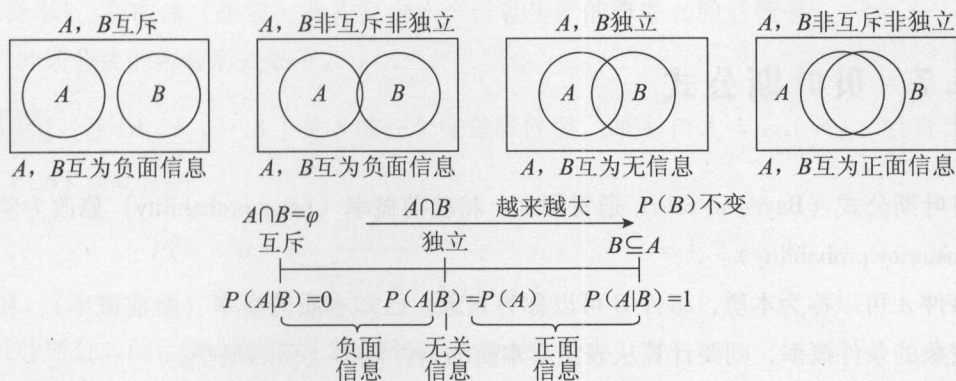


图 5-13 两事件的互斥、独立和信息的关系

例题 5.19 独立事件与互斥事件。(解答见网络资源)

定义 若事件 A_1, A_2, \dots, A_n 为 n 个事件, 对于任何其中 k 个事件:

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \times P(A_{i_2}) \times \dots \times P(A_{i_k}), \quad k = 2, \dots, n$$

则称事件 A_1, A_2, \dots, A_n 为相互独立事件 (mutually independent)。

若事件 A, B 和 C 为相互独立事件, 则

$$P(A \cap B) = P(A)P(B),$$

$$P(A \cap C) = P(A)P(C), \quad P(B \cap C) = P(B)P(C),$$

$$P(A \cap B \cap C) = P(A)P(B)P(C), \quad P(A|B \cap C) = P(A)$$

定义 若事件 A_1, A_2, \dots, A_n 为 n 个事件, 对于任何其中两个事件:

$$P(A_i \cap A_j) = P(A_i) \times P(A_j), \quad i \neq j$$

则称事件 A_1, A_2, \dots, A_n 为两两独立事件 (pairwise mutually independent)。

相互独立事件, 一定是两两独立。但是两两独立事件, 不一定是相互独立。

事件 A_1, A_2, \dots, A_n 有 4 种情况: ①互斥事件; ②非两两独立且非互斥事件; ③相互独立事件; ④两两独立但非相互独立事件, 如图 5-14 所示。

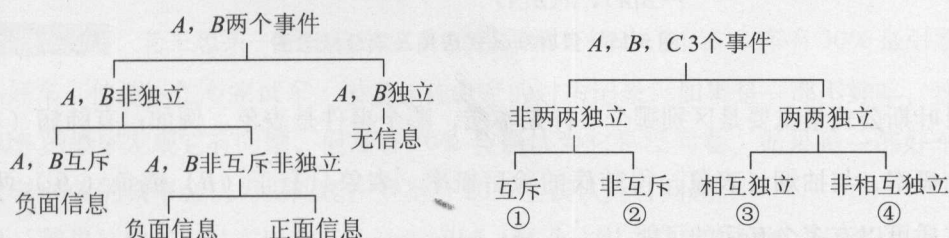


图 5-14 事件的独立与互斥

5.7 贝叶斯公式

贝叶斯公式 (Bayesian rule) 通常用在, 将验前概率 (prior probability) 修改为验后概率 (posterior probability)。

事件 A 可以称为本质, 事件 B 可以称为表象。已知本质的概率 (验前概率), 和从本质看表象的条件概率, 则要计算从表象看本质的条件概率 (验后概率)。

例如: 有病 (癌症) 是本质, 检验结果阳性是表象, 已知癌症的概率 (验前概率), 和癌症的检验结果是阳性的条件概率, 则要计算检验结果阳性, 会是癌症的概率 (验后概率)。

定理 若事件 A 与 B 为同一试验的事件, 则

$$\begin{aligned} P(B) &= P(A \cap B) + P(\bar{A} \cap B) = P(A \cap B) + P(B - A) \\ P(B) &= P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) \end{aligned}$$

定理 全概率公式 (total probability theorem): 若事件 A_1, A_2, \dots, A_n 为一个完备事件组, 则

$$P(B) = \sum_{i=1}^n P(B \cap A_i) = \sum_{i=1}^n P(B|A_i)P(A_i)$$

定理 贝叶斯公式: 若事件 A 与 B 为同一样本空间的事件, 且 $P(B) \neq 0$, 则

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$

贝叶斯公式的应用如图 5-15 所示。

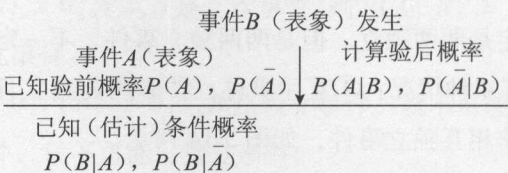


图 5-15 贝叶斯公式应用及集合概念图

贝叶斯公式最重要是区别哪个事件是本质, 哪个事件是表象。例如: 有肺病 (本质) 的验前概率, 与抽烟 (表象) 得肺病的验后概率。表象只有正 (B) 或负 (\bar{B}) 两个结果, 本质可以有多个互斥的可能 $\{A_1, A_2, \dots, A_k\}$ 。

例如: 有 3 台生产设备 (本质) 的生产比率 (验前概率), 每台生产次品 (表象) 率

(条件概率), 则次品(表象)是由某台生产设备生产的概率(验后概率)。

贝叶斯公式的一般形式如下。

定理 若 $\{A_1, A_2, \dots, A_k\}$ 为 S 的一个完备事件组, 即 $A_i \cap A_j = \varphi, i \neq j$, 且 B 为另一事件, $P(B) \neq 0$, 则

$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{j=1}^k P(B | A_j)P(A_j)} \quad i = 1, 2, \dots, k$$

贝叶斯公式的计算步骤如表 5-6 和表 5-7 所示。

表 5-6 贝叶斯公式的计算步骤 (A_1, A_2, B, \bar{B})

列联表		表象		验前概率 已知		$P(\bar{B} A_i)$
		B	\bar{B}	$P(A_i)$	$P(B A_i)$	
本 质	A_1	⑦ = ① × ③	⑨ = ① × ⑤	①	③	⑤ = 1 - ③
	A_2	⑧ = ② × ④	⑩ = ② × ⑥	②	④	⑥ = 1 - ④
$P(B_i)$		⑪ = ⑦ + ⑧	⑫ = ⑨ + ⑩	1		
$P(A_1 B_i)$		⑬ = ⑦/⑪	⑮ = ⑨/⑫	} 验后 概率		
$P(A_2 B_i)$		⑭ = ⑧/⑪	⑯ = ⑩/⑫			

表 5-7 贝叶斯公式的计算步骤 ($A_1, A_2, A_3, B, \bar{B}$)

列联表		表象		验前概率 已知		$P(\bar{B} A_i)$
		B	\bar{B}	$P(A_i)$	$P(B A_i)$	
本 质	A_1	⑩ = ① × ④	⑬ = ① × ⑦	①	④	⑦ = 1 - ④
	A_2	⑪ = ② × ⑤	⑭ = ② × ⑧	②	⑤	⑧ = 1 - ⑤
	A_3	⑫ = ③ × ⑥	⑮ = ③ × ⑨	③	⑥	⑨ = 1 - ⑥
$P(B_i)$		⑯ = ⑩ + ⑪ + ⑫	⑰ = ⑬ + ⑭ + ⑮	1		
$P(A_1 B_i)$		⑱ = ⑩/⑯	㉑ = ⑬/⑰	} 验后 概率		
$P(A_2 B_i)$		㉒ = ⑪/⑯	㉓ = ⑭/⑰			
$P(A_3 B_i)$		㉔ = ⑫/⑯	㉕ = ⑮/⑰			

可以从表 5-6 和表 5-7 检查 A_i 和 B_i , 是独立、正面信息或负面信息。

例题 5.20 老王想买一辆中古的 C 车, 市场调查统计, 中古 C 车有 30% 是引擎有毛病的不好车。他请一位专家试车。根据这位专家的去记录, 如果是一部不好车, 这位专家有 90% 的概率发现它有问题, 但是有 10% 会误认为它是没问题。如果是一部好车, 这位专家有 80% 的概率会认为没问题, 但是有 20% 会误认为有问题。

- (1) 如果这位专家试车以后, 认为没问题, 那么这车子是好车的概率是多少?
- (2) 如果这位专家试车以后, 认为有问题, 那么这车子是坏车的概率是多少?

买到好车的事前概率，与试车以后没有问题而是好车的事后概率如图 5-16 所示。

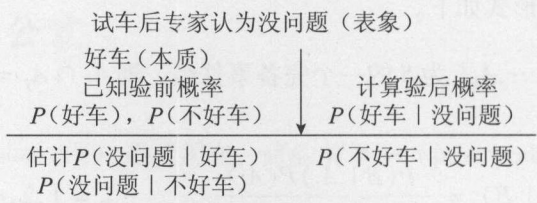


图 5-16 各种情况的概率

解答：令 A = 这车子是好车， \bar{A} = 这车子不是好车。

令 B = 专家认为没问题， \bar{B} = 专家认为有毛病。

事前（试车前）概率

$P(A) = 0.7, P(\bar{A}) = 0.3$

$P(B | A) = 0.8, P(\bar{B} | A) = 0.2, P(B | \bar{A}) = 0.1, P(\bar{B} | \bar{A}) = 0.9$

全概率公式

$P(B) = P(B | A)P(A) + P(B | \bar{A})P(\bar{A}) = (0.8)(0.7) + (0.1)(0.3) = 0.59$

因为 $P(B | A) \geq P(B)$ ，所以 A 和 B 互为正面信息。

答案：

(1) $P(A | B) = \frac{P(B | A)P(A)}{P(B | A)P(A) + P(B | \bar{A})P(\bar{A})} = \frac{(0.8)(0.7)}{(0.8)(0.7) + (0.1)(0.3)} = 0.95$

(2) $P(\bar{A} | \bar{B}) = \frac{P(\bar{B} | \bar{A})P(\bar{A})}{P(\bar{B} | \bar{A})P(\bar{A}) + P(\bar{B} | A)P(A)} = \frac{(0.9)(0.3)}{(0.9)(0.3) + (0.2)(0.7)}$

$= 0.66$

结论：如果专家认为没问题，则是好车的概率，从 0.7 升到 0.95（正面信息）；如果专家认为有毛病，则不好车的概率，从 0.3 升到 0.66（正面信息），如表 5-8 所示。

表 5-8 好车与不好车的概率

专家认为 实际是	没问题 B_1	有问题 B_2	验前概率 已知		$P(B_2 A_i)$
			$P(A_i)$	$P(B_1 A_i)$	
好车 A_1	0.56	0.14	0.7	0.8	0.2
不好车 A_2	0.03	0.27	0.3	0.1	0.9
$P(B_i)$	0.59	0.41	1		
$P(A_1 B_i)$	0.95	0.34			
$P(A_2 B_i)$	0.05	0.66			

上述结果，请对照图 5-7（a）的情况。

例题 5.21 (见网络资源)

5.8 中文统计应用

5.8.1 贝叶斯定理 (例题 5.20)

执行“贝叶斯定理”的操作示意图和结果分别如图 5-17 和图 5-18 所示。

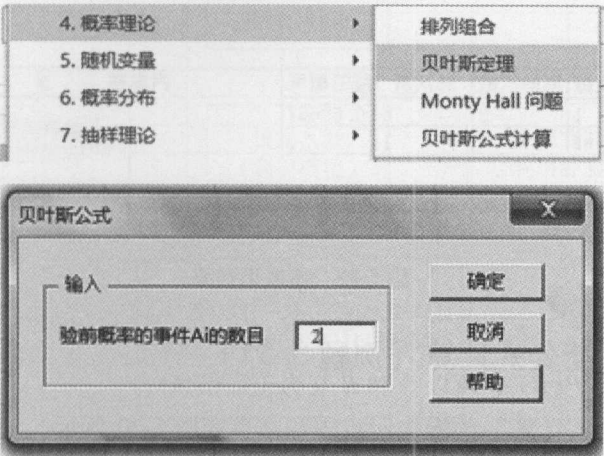


图 5-17 执行“贝叶斯定理”的操作示意图

	A	B	C	D	E
1	贝叶斯公式				
2	事件数目	2		重新计算	
3					
4	事件	验前概率	条件概率	交集概率	验后概率
5	Ai	P(Ai)	P(B Ai)	P(Ai ∩ B)	P(Ai B)
6	A1	0.7	0.8	0.56	0.949153
7	A2	0.3	0.1	0.03	0.050847
8	Total	1	P(B)=	0.59	

图 5-18 执行“贝叶斯定理”的结果

5.8.2 Monty Hall 问题 (例题 5.17)

换门游戏：请选择一个门→选择要不要换→结果是赢或输→累积概率。
模拟：“换门”100 次模拟的累积概率或“不换”100 次模拟的累积概率。
模拟示意图如图 5-19 所示。

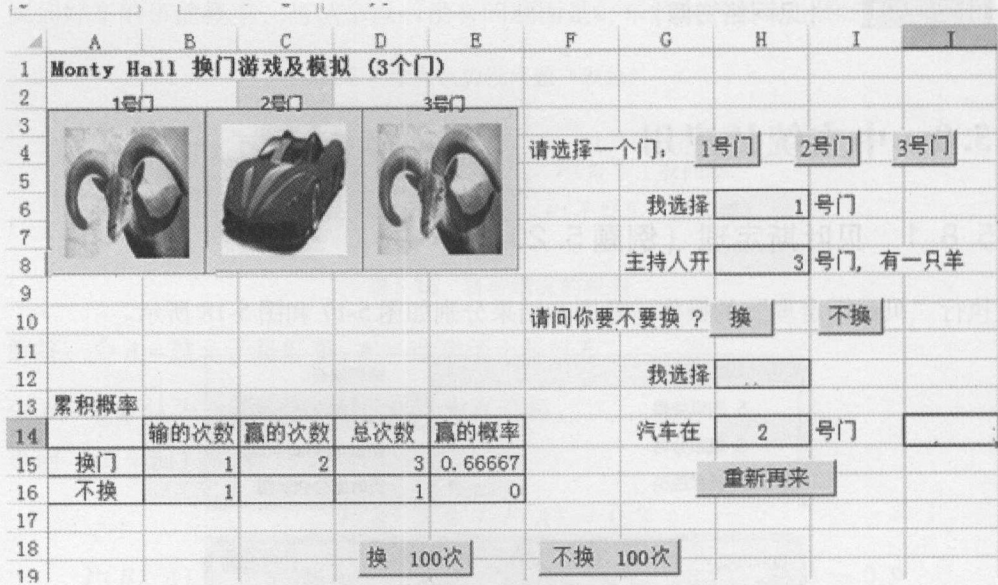


图 5-19 模拟示意图

5.8.3 贝叶斯公式计算 (例题 5.21)

执行“贝叶斯公式计算”的操作示意图和结果如图 5-20 所示。



图 5-20 执行“贝叶斯公式计算”的操作示意图和结果

5.9 本章流程图

本章流程图如图 5-21 所示。

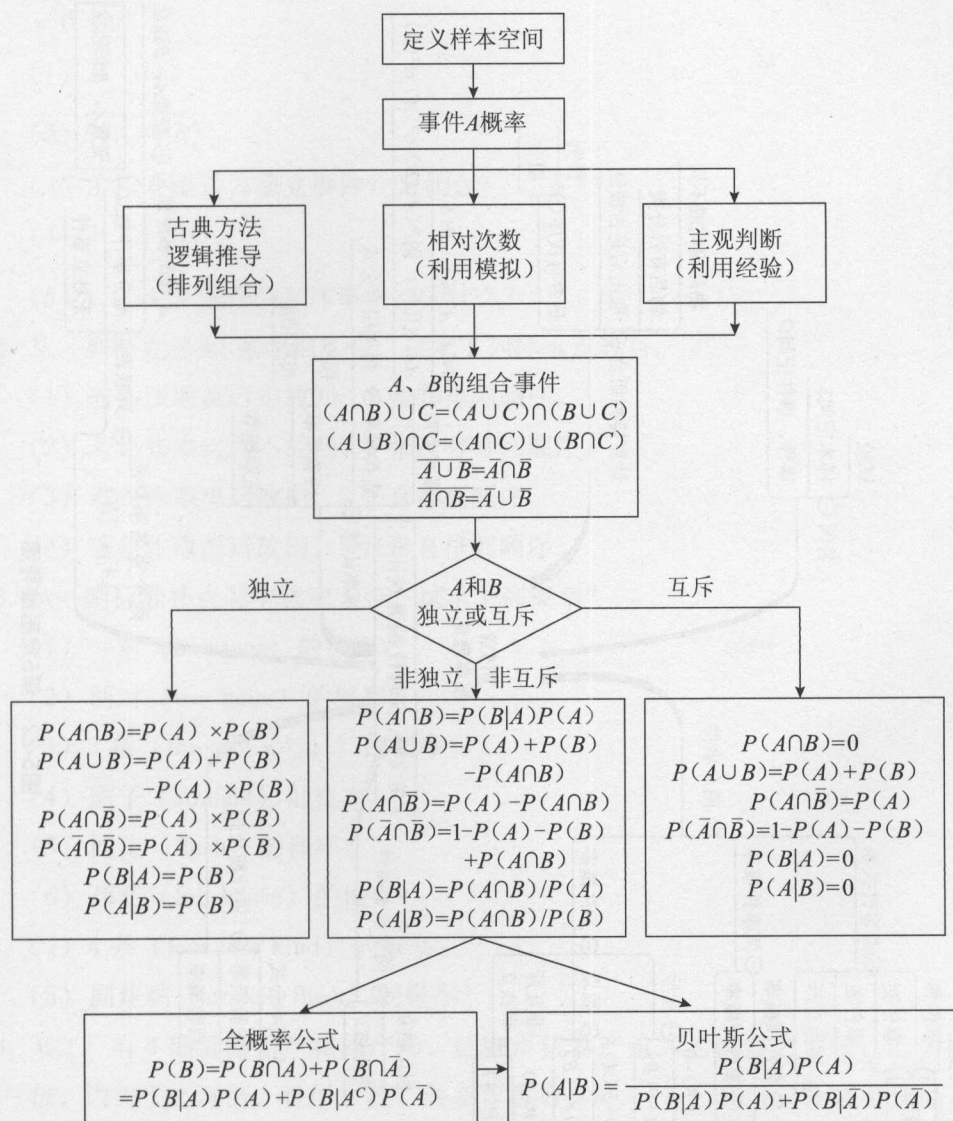


图 5-21 第 5 章流程图

5.10 本章思维导图

本章思维导图如图 5-22 所示。

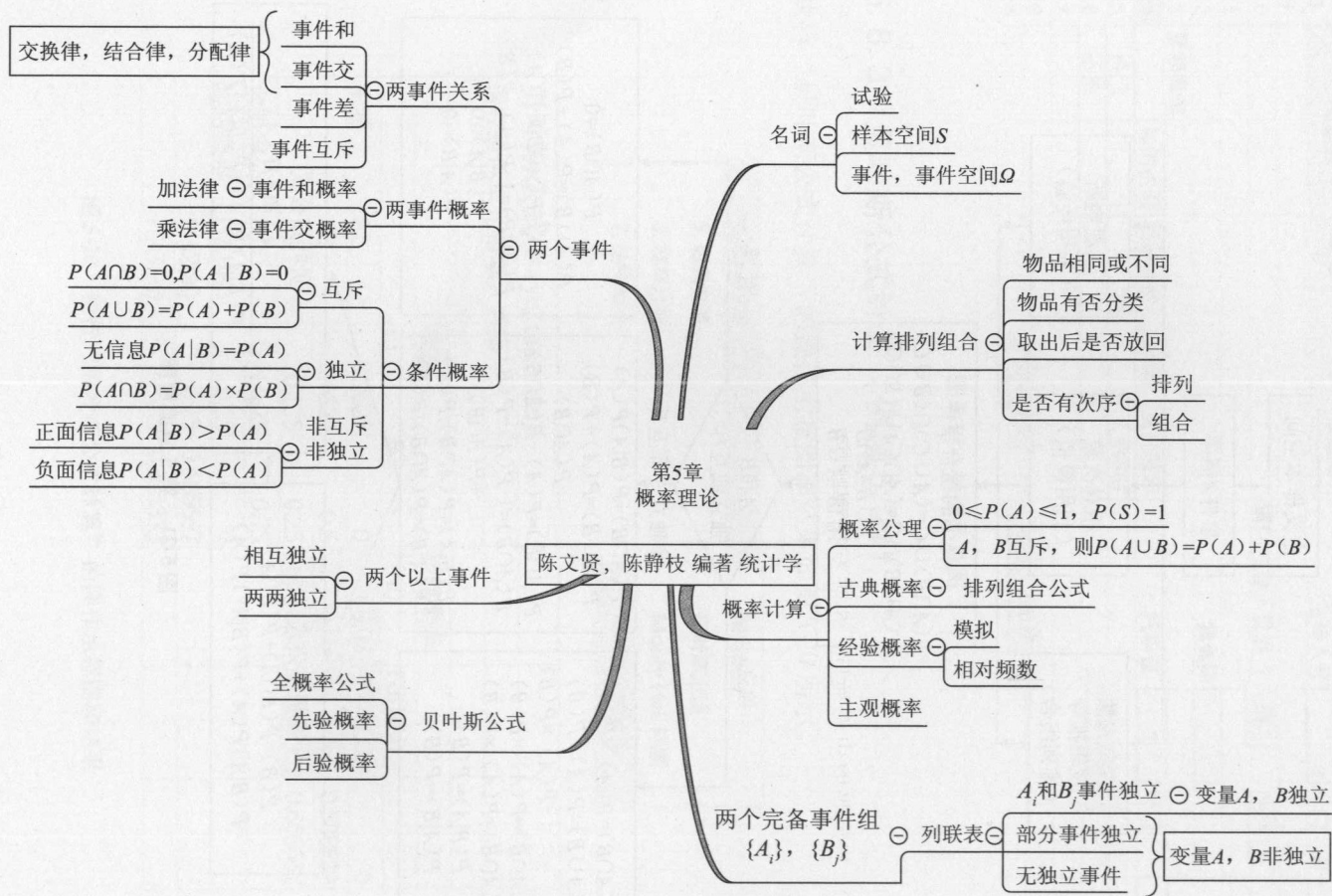


图5-22 第5章思维导图

习 题

- 令 A, B 为样本空间 S 中的两个事件, 若 $P(B) = 0.3, P(A \cap B) = 0.12, P(\bar{A} \cap \bar{B}) = 0.42$
 - (1) $P(A)$ 。
 - (2) $P(A \cap \bar{B})$ 。
 - (3) A 和 B 是否为独立事件? 为什么?
 - (4) $P(A|B)$ 。
 - (5) A, B, \bar{A}, \bar{B} 是否为互斥事件? 为什么?
- 从一副扑克牌 52 张中取 5 张, 试求下列样本点数目。
 - (1) 若每张取出后不放回, 5 张有排列顺序。
 - (2) 若每张取出后不放回, 5 张没有排列顺序。
 - (3) 若每张取出后放回, 5 张有排列顺序。
 - (4) 若每张取出后放回, 5 张没有排列顺序。
- 从一副标准扑克牌中选取 5 张, 试求下列概率?
 - (1) 一对 (one pair) 的概率?
 - (2) 两对 (two pairs) 的概率?
 - (3) 3 条 (three of a kind) 的概率?
 - (4) 顺子 (straight) 的概率?
 - (5) 同花 (flush) 的概率?
 - (6) 葫芦 (full house) 的概率?
 - (7) 4 条 (four of a kind) 的概率?
 - (8) 同花顺 (straight flush) 的概率?
- 某工厂有 4 部机器生产相同产品, 已知甲机器产量为乙机器之两倍, 乙为丙之两倍, 丙为丁之两倍, 且甲、乙、丙及丁各有 5%, 4%, 3%, 2% 之不良品, 今随机抽取一件, 发现其为不良品, 试问此一产品为甲机器所生产之概率为若干?
- 某甲在其上班公司之一份年度报告中称: 该公司 60 名新进人员中, 会游泳的 35 人, 会开汽车 24 人, 会使用计算机的 25 人, 会游泳又会开车的 12 人, 会游泳又会使用计算机的 10 人, 会开车又会使用计算机的 8 人, 三者皆会的 7 人。这份报

告呈上后，某甲即被上司判定工作不力，为何？

6. 某公司鼓励员工空闲时多做运动，为了增购运动器材，调查所有员工对于运动的嗜好，有 60% 的员工喜欢打乒乓球，有 50% 的员工喜欢打羽毛球，两者都喜欢的占 30%，设 E 为员工喜欢打乒乓球的事件， F 为员工喜欢打羽毛球的事件。

(1) 请问下列概率值为多少？

① $P(E)$ ；

② $P(F)$ ；

③ $P(E \cap F)$ ；

④ $P(E \cup F)$ ；

⑤ $P(F|E)$ ；

⑥ $P(E|F)$ 。

(2) 员工打乒乓球的嗜好与打羽毛球的嗜好是否独立？

7. 某一数学系有 12 个成员，5 位为高年级，7 位为低年级，如果要从其中任意选 4 人参加会议，试求 4 人当中高年级多于低年级的概率。

其他习题请下载。



第6章

随机变量

随：元亨利贞，无咎。

随，《说文》：“从也。”《广雅·释诂》：“随，顺也。”

——《易经第十七卦随卦》

山水之法，在乎随机应变。

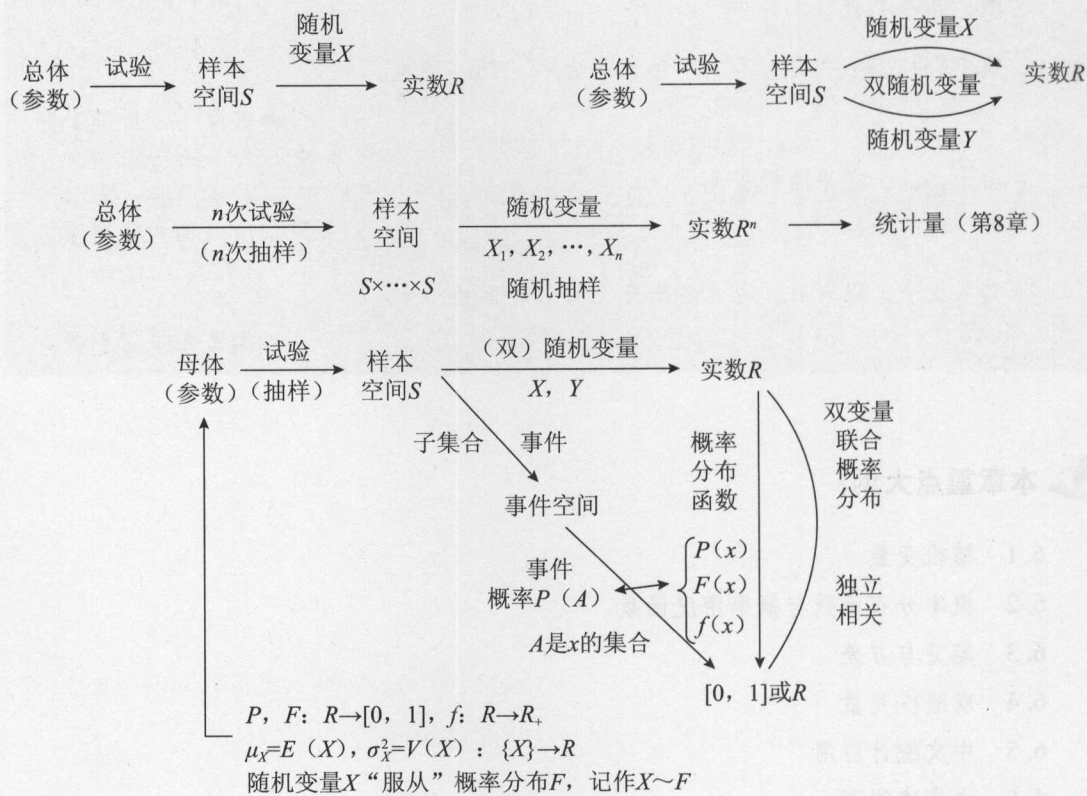
——元·陶宗仪《南村辍耕录》

兵以正合，以奇胜，善之者出奇无穷，奇正还相生。

——司马迁《史记》

本章重点大纲：

- 6.1 随机变量
- 6.2 概率分布函数与概率密度函数
- 6.3 期望与方差
- 6.4 双随机变量
- 6.5 中文统计应用
- 6.6 本章流程图
- 6.7 本章思维导图



本章概念图

6.1 随机变量

样本空间的样本点，通常是用文字（如：员工姓名），符号（如：骰子），或数字（如：产品编号）等来表示。但是，在处理事件的概率上，用文字或符号表达样本点的集合（即事件），就比较麻烦。同时，我们对数值的结果（如平均数等）比较有兴趣，所以要定义一个将样本空间对应到实数的函数，称作随机变量。

每个随机变量会有对应的概率函数、期望（均值）、方差、偏态系数、峰态系数等。

随机变量是将“样本点”转换成“实数”，配合概率的运算，利用抽样的随机变量，可计算或推断均值、方差等总体的特征（第6, 7, 8章）。两个随机变量，代表一个总体的两个变量，或两个总体的同一个变量，可以做因果关系的差异、关系和独立的推断（第9, 10, 11, 12章）。

定义 随机变量（random variable），是一个将样本空间 S 对应到实数 R 的函数，通常记作 X 。从本章开始，我们用英文大写字母如 X 、 Y 等表示随机变量；用英文小写字母如 x 、 y 等表示实数，如 $X:S \rightarrow R$ 。

一个样本空间的随机变量，是将样本点（例如：产品），配合要统计的主题（例如：产品质量，或产品是否合格：合格为1、不合格为0），对应到实数值。

所以，同一样本空间，不同的统计主题，有不同的随机变量。

随机变量 X 的值域（range），记作 R_X 。

$$R_X = \{X(E_i) \mid E_i \in S\}$$

定义 若随机变量 X 的值域 R_X 是有限或可数的无限（countable infinite，如1, 2, 3, ...），则称 X 为离散型或间断型（discrete）随机变量。

定义 若随机变量 X 的值域 R_X 是包括一个实数区间（interval），如： $[0, 6]$ ，或 $[0, \infty)$ 等，则称 X 为连续型（continuous）随机变量。

R_X 除了离散型（间断型）或连续型，还有衡量尺度：比率、区间、顺序、名目，另外还有单位（长度、质量、金钱、分数、等级、个数等）。因此，随机变量可以做数学运算、排序、归类、函数等。以下是随机变量的举例。

1) 3个小孩出生的性别，令 B 代表男孩， G 代表女孩，随机变量 X （样本点3个婴儿）= 男孩的数目，则

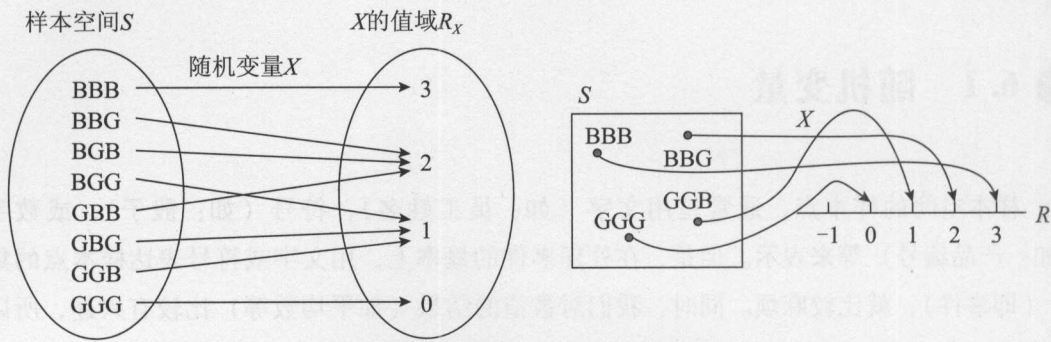


图 6-1 样本空间、随机变量与值域

2) 掷两个骰子，我们知道有 36 个样本点。定义样本点 E_{ij} 为第 1 个骰子出现 i 点，第 2 个骰子出现 j 点， $i, j = 1, 2, 3, 4, 5, 6$ 。

如果要计算两个骰子的点数和的概率，则随机变量 X ：

$$X(E_{ij}) = i + j, R_X = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

如果要计算两个骰子的点数差的概率，则随机变量 Y ：

$$Y(E_{ij}) = |i - j|, R_Y = \{0, 1, 2, 3, 4, 5\}$$

3) 掷两个硬币，我们知道有 4 个样本点。定义 H 为人头，T 为反面，样本空间为 $\{HH, HT, TH, TT\}$ 。如果要计算出现人头的数目的概率，则随机变量 X ：

$$X(HH) = 2, X(HT) = 1, X(TH) = 1, X(TT) = 0, R_X = \{0, 1, 2\}$$

4) 100 个学生，代表 100 个样本点。定义样本点 A_i 为第 i 个学生， $i = 1, 2, \dots, 100$ 。如果要计算男女学生性别比例，则随机变量 X ：

$$X(A_i) = \text{第 } i \text{ 个学生的性别}, R_X = \{1, 2\}$$

若 A_i 是男性，则 $X(A_i) = 1$ ；若 A_i 是女性，则 $X(A_i) = 2$ 。

如果要计算学生分数的分布，则随机变量 X ：

$$X(A_i) = \text{第 } i \text{ 个学生的总成绩}, R_X = [0, 100] \text{ (0 分到 100 分)}$$

如果要计算学生身高的平均数，则随机变量 X ：

$$X(A_i) = \text{第 } i \text{ 个学生的身高}, R_X = [150, 200] \text{ (150cm 到 200cm)}$$

5) 100 件产品，已知有 20 件故障品。抽出 2 件产品，则样本点有 $C_2^{100} = 4950$ 个。如果要计算出现故障品的概率，则随机变量 X ：

$$X = \text{两件产品中故障的数目}, R_X = \{0, 1, 2\}$$

6.2 概率分布函数与概率密度函数

因为样本空间的事件，有一个概率函数（见 6.4 概率的计算），所以在随机变量的值域（即实数），也可以定义一个概率函数如下：

$$P: R_X \rightarrow [0, 1]$$

随机变量 X 的值域是概率函数 P 的定义域，即

$$P(r) = P(X = r) = P(\{\text{事件点 } E_i \mid X(E_i) = r\})$$

同理

$$P(X > r) = P(\{\text{事件点 } E_i \mid X(E_i) > r\})$$

定义 若 X 是离散型随机变量，则 X 的概率分布函数（probability distribution function，简称 p. d. f.），或称为 X 的概率质量函数（probability mass function，简称 p. m. f.），记作 $P(X=x)$ ，简记 $P(x)$ ， $P(x) = P(X=x) = P(\{\text{事件点 } E_i \mid X(E_i) = x\})$ ，满足下列条件：

$$(1) 0 \leq P(x) = P(X=x) \leq 1$$

$$(2) \sum_x P(x) = 1$$

$$(3) P(c \leq X \leq d) = \sum_{x=c}^d P(x)$$

定义 若 X 是连续型随机变量，其值域为区间 $[a, b]$ ，则 X 的概率密度函数（probability density function，简称 p. d. f.），记作 $f_X(x)$ ，或简记 $f(x)$ ，满足下列条件：

$$(1) f(x) \geq 0, \quad a \leq x \leq b \quad (\text{注意: } f(x) \text{ 有可能大于 } 1)$$

$$(2) \int_a^b f(x) dx = 1$$

$$(3) P(c \leq x \leq d) = \int_c^d f(x) dx, \quad a \leq c < d \leq b$$

注意：以上虽然都简称 p. d. f.，但是要分辨其是离散型或连续型。

若 X 是连续型随机变量，则任何 $c \in R$ ：

$$P(X=c) = \int_c^c f(x) dx = 0$$

所以，连续型随机变量在任何一点的概率为 0。

但是只要加上一点距离，则 $f(c)$ 还是代表随机变量在该点附近的概率。概率密度好

比人口密度, 在任何一点的概率 (人口数) 为 0, 但是要有一段区间 (一块面积) 才有概率 (人口), 即

$$P\left[\left(c - \frac{\varepsilon}{2}\right) \leq x \leq \left(c + \frac{\varepsilon}{2}\right)\right] = \int_{c-\frac{\varepsilon}{2}}^{c+\frac{\varepsilon}{2}} f(x) dx \approx \varepsilon f(c)$$

若 X 是连续型随机变量, 则任何 $c, d \in R, c < d$:

$$P(c \leq x \leq d) = P(c < x \leq d) = P(c \leq x < d) = P(c < x < d)$$

但是离散型随机变量, 上述式子, 不一定成立。

正态分布的概率密度图如图 6-2 所示。

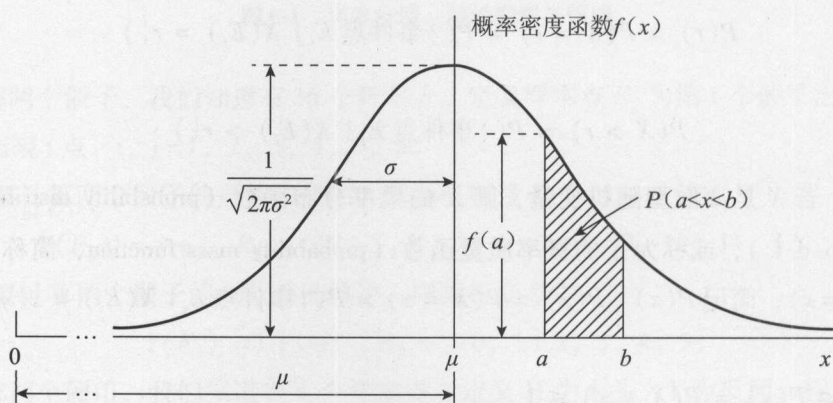


图 6-2 正态分布的概率密度图

定义 若 X 是离散型随机变量, 则 X 的累积分布函数 (cumulative distribution function, 简称 c. d. f.) 为 $F(X=x)$, 简记 $F(x)$, 其公式为

$$F(x) = F(X=x) = P(X \leq x) = \sum_{i \leq x} P(X=i)$$

定义 若 X 是连续型随机变量, 则 X 的累积分布函数 (cumulative distribution function, 简称 c. d. f.) 为 $F(X=x)$, 简记 $F(x)$, 其公式为

$$F(x) = F(X=x) = P(X \leq x) = \int_{-\infty}^x f(y) dy$$

$$\forall a < b \Rightarrow F(a) \leq F(b)$$

$$P(X > x) = 1 - P(X \leq x) = 1 - F(x)$$

性质 若 X 是离散型随机变量, 则 $F(x)$ 是阶梯形上升函数, 如图 6-3 所示。

性质 若 X 是连续型随机变量, 则 $F(x)$ 是连续形上升函数, 如图 6-4 所示。 $\forall a < b \Rightarrow F(a) \leq F(b)$ 。

$$\frac{d}{dx}F(x) = f(x)$$

$$P(X \geq x) = P(X > x) = 1 - P(X \leq x) = 1 - F(x)$$

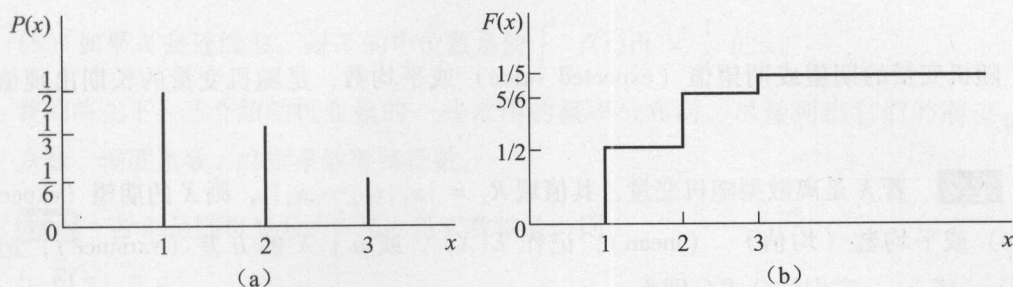


图 6-3 离散型的概率分布函数与累积分布函数

(a) 概率分布函数; (b) 累积分布函数

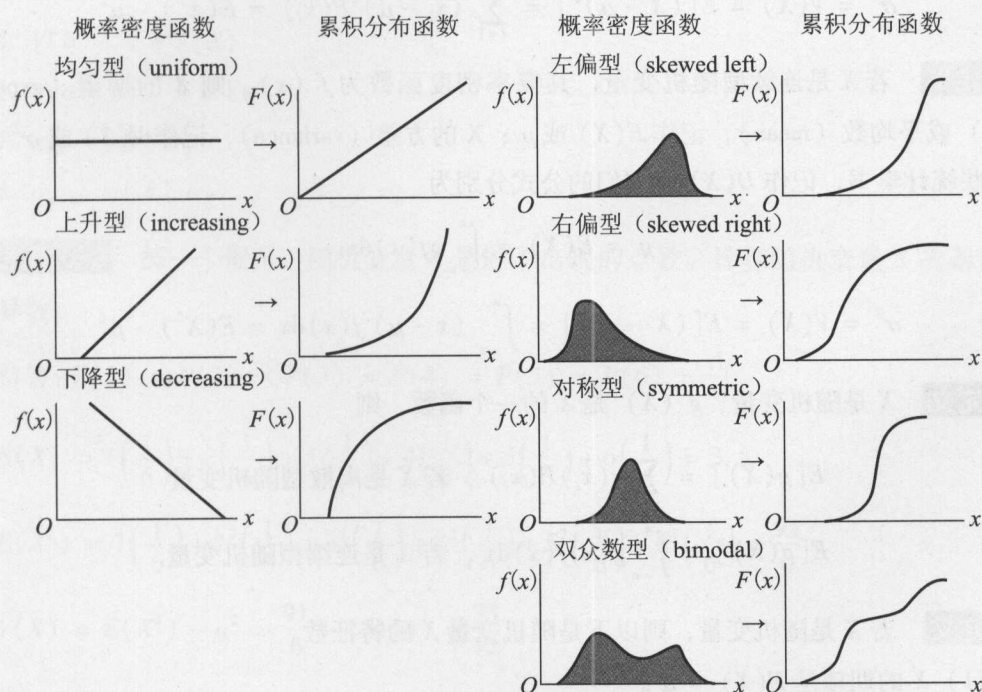


图 6-4 几种连续型的概率密度函数和累积分布函数

例题 6.1 随机变量。(见网络资源)

例题 6.2 累积分布函数。(见网络资源)

6.3 期望与方差

随机变量的期望或期望值 (expected value) 或平均数, 是随机变量的长期出现值的平均。

定义 若 X 是离散型随机变量, 其值域 $R_X = \{x_1, x_2, \dots, x_k\}$, 则 X 的期望 (expected value) 或平均数 (均值) (mean), 记作 $E(X)$, 或 μ ; X 的方差 (variance), 记作 $V(X)$, 或 σ^2 。它们的公式分别为

$$\mu = E(X) = \sum_{i=1}^k x_i P(x_i)$$

$$\sigma^2 = V(X) = E[(X - \mu)^2] = \sum_{i=1}^k (x_i - \mu)^2 P(x_i) = E(X^2) - \mu^2$$

定义 若 X 是连续型随机变量, 其概率密度函数为 $f(x)$, 则 X 的期望 (expected value) 或平均数 (mean), 记作 $E(X)$ 或 μ ; X 的方差 (variance), 记作 $V(X)$ 或 σ^2 , 国内有些统计学书, 记作 $D(X)$ 。它们的公式分别为

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x) dx$$

$$\sigma^2 = V(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = E(X^2) - \mu^2$$

定义 X 是随机变量, $g(X)$ 是 X 的一个函数, 则

$$E[g(X)] = \sum_{i=1}^k g(x_i) P(x_i), \text{ 若 } X \text{ 是离散型随机变量}$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x) dx, \text{ 若 } X \text{ 是连续型随机变量}$$

定义 若 X 是随机变量, 则以下是随机变量 X 的特征数。

- (1) X 的期望是 $E(X) = \mu$ 。
- (2) X 的方差是 $V(X) = \sigma^2 = \mu_2$ 。
- (3) X 的标准差是 $\sigma_X = \sqrt{\sigma^2}$ 。
- (4) X 的三阶中心距是 $M_3(X) = E[(X - \mu)^3] = \mu_3$ 。
- (5) X 的四阶中心距是 $M_4(X) = E[(X - \mu)^4] = \mu_4$ 。
- (6) X 的偏度系数是 $S(X) = \frac{M_3(X)}{\sigma^3}$ 。

(7) X 的峰度系数是 $K(X) = \frac{M_4(X)}{\sigma^4}$ 。

(8) X 的众数是使 $P(x_i)$ 或 $f(x)$ 最大的 x_i 或 x 。

(9) 如果 X 是连续型, 则 X 的中位数是使 $\int_{-\infty}^x f(t) dt = \frac{1}{2}$ 的 x 。

我们将在下一章介绍随机变量的一些常用的概率分布时, 尽量列出它们的期望、方差、众数、偏度系数、峰度系数等特征数。

定理 若 X 是随机变量, a 与 b 是实数常数, 则

1. $E(a) = a$
2. $E(bX) = bE(X)$
3. $E(a + bX) = a + bE(X)$
4. $V(a) = 0$
5. $V(bX) = b^2V(X)$
6. $V(a + bX) = b^2V(X)$
7. $\sigma_{bX} = |b| \sigma_X$
8. $\sigma_{a+bX} = |b| \sigma_X$

例题 6.3 掷一个骰子, 随机变量 X 是骰子出现的点数。计算随机变量 X 的期望值与变异数。

解答: $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$

$$E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3.5$$

$$E(X^2) = 1\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right) = \frac{91}{6}$$

$$V(X) = E(X^2) - \mu^2 = \frac{91}{6} - (3.5)^2 = \frac{35}{12}$$

例题 6.4 竞价的期望值。(解答见网络资源)

例题 6.5 赌场优势

赌场优势 (House advantage, House Edge) 是赌场的每一项赌博游戏, 玩家押注一单位, 其“玩家期望的负数”, 以百分比计算。赌场优势 = 玩家期望的负数 = 赌场的期望。通常玩家的期望是负的, 负负得正, 赌场优势是正的, 赌场优势越大, 对玩家越不公平。长期来看, 玩家永远是输的, 赌场永远是赢的。请计算赌场里, 轮盘 (Roulette) 和百家

乐 (Baccarat) 的赌场优势。

解答：美式轮盘 (请见第 5 章习题 29 题) 有 38 个格子 (1~36 及 0, 00)，有 9 种押注策略，押注 1 个号码到押注 18 个号码，每个押注策略是一个随机变量。如果只押 1 个号码，则赔 35 倍。玩家的期望是：

$$35 \times \frac{1}{38} + (-1) \times \frac{37}{38} = -\frac{2}{38} = -0.0526 = -5.26\%$$

所以，赌场优势是 5.26%。除了押 5 个号码 (赔 6 倍) 的赌场优势是 7.89%，其他 8 种押注策略的赌场优势都是 5.26%。但是方差就不相同，押注号码越少，赔的倍数越大，方差 (风险) 也越大。

百家乐使用 6 或 8 副扑克牌，是一个比较复杂的规则，有 3 种押注策略：庄家 (banker)、闲家 (player)、和局 (tie)，其概率计算无法用逻辑推导，而用模拟方法。用 8 副牌：庄家赢概率 45.86%、闲家赢概率 44.62%、和局赢概率 9.52%。押庄家赢赔 0.95 倍、押闲家赢赔 1 倍、押和局赢赔 8 倍。押庄或押闲，出现和局，不输不赢。所以，押庄家赢的赌场优势是 1.053%：

$$0.95 \times 0.4586 + (-1) \times 0.4462 + 0 \times 0.0952 = -0.01053 = -1.053\%$$

我们将常见的赌博游戏的赌场优势，如表 6-1 所示。

百家乐的 8 副 (扑克) 牌比 6 副牌或 1 副牌，提高押庄家赢的赌场优势，但是降低押闲家与押和局的赌场优势。21 点的基本策略是一个表，可以上网查。21 点的 6 副牌比 1 副牌，赌场优势大，但是使玩家可以“算牌”，当剩下来的牌，大牌的数目多时，庄家爆牌概率高，玩家就下大赌注，可以使玩家的期望更大，不只是 1%。

赌博产业的三大原则。

- (1) 诚实：出象 (牌) 的“随机性”，概率问题，双方没有诈欺或出牌给特定人。
- (2) 公平：游戏规则 → 赌场优势 (越小越公平) → 统计学 → 赌场数学。
- (3) 履约：双方依约付 (赔) 款，没有赖账，愿赌服输。

保险产业的三大原则。

- (1) 诚实：保户出事的“随机性”，纯属意外 (意料之外)，保户没有诈欺。
- (2) 公平：保费赔偿的计算 → 精算学 (利用很多统计学) → 保险数学。
- (3) 履约：双方依约付 (赔) 款。

表 6-1 赌博游戏的赌场优势

赌博游戏	押注策略	赌场优势	
百家乐 (Baccarat)		8 副牌	1 副牌
	押庄家 (赔 0.95)	1.05%	1.01%
	押闲家 (赔 1)	1.24%	1.28%
	押和局 (赔 8)	14.32%	15.76%
21 点 (Blackjack)		6 副牌	1 副牌
	基本策略	0.5%	0.2%
	平均玩家	2%	<2%
	算牌玩家	-1.0%	
轮盘 (Roulette)		有两个 0	只有 1 个 0
	除了押 5 个号码	5.26%	2.76%
	押 5 个号码	7.89%	
Keno		27%	
牌九 (Pai Gow)		2.54% ~ 2.84%	
吃角子老虎 (Slot)		4% ~ 15%	
计算机扑克 (Poker)		2.32% ~ 3.37%	

6.4 双随机变量

有时候，统计学要研究两个变量之间的关系。例如：一个人“身高”与“体重”的关系；一个公司一季“广告支出”与“销售额”的关系。企业“信息科技投资”与“获利率”的关系。

当一个试验，包含两个以上随机变量，其统计分析，称之为多变量分析 (multivariate analysis)。双随机变量的两个随机变量，要有相同的样本空间。

例如：抽出一个人 (这是一个试验)，包括“身高”与“体重”两个随机变量；抽出一个公司一季财务报表 (这是一个试验)，有“季广告支出”与“季销售额”两个随机变量；掷两个骰子，有“点数和”与“点数差”两个随机变量。下面我们介绍双随机变量 (bivariate)。

定义 若 X 和 Y 是离散型随机变量，则 X, Y 的联合概率分布函数 (joint probability distribution function) 为 $P(X = x, Y = y)$ ，简记 $P(x, y)$ ，可以图 6-5 (a) 表示，满足：

(1) $P(x,y) = P(X = x,Y = y) = P(\{X = x\} \cap \{Y = y\}) , \quad 0 \leqslant P(x,y) \leqslant 1$

(2) $\sum_x \sum_y P(x,y) = 1$

(3) $P(a \leqslant X \leqslant b,c \leqslant Y \leqslant d) = \sum_{x=a}^b \sum_{y=c}^d P(x,y)$

若 X, Y 是离散型随机变量, 则 X, Y 的联合概率分布函数, 可以用矩阵表 (二分类列联表) 来表示。

例题 6.6 如果有: 3 个红球, 4 个白球, 5 个黄球。随机抽出 3 个球, 抽出后不放回。令 X =抽出 3 个球中红球的数目; Y =抽出 3 个球中白球的数目。计算联合概率分布函数 $P(i,j) = P(X = i,Y = j)$ 。

解答: $P(0,0) = C_3^5/C_3^{12} = 10/220$, $P(0,1) = C_1^4C_2^5/C_3^{12} = 40/220$
 $P(0,2) = C_2^4C_1^5/C_3^{12} = 30/220$, $P(0,3) = C_3^4/C_3^{12} = 4/220$
 $P(1,0) = C_1^3C_2^5/C_3^{12} = 30/220$, $P(1,1) = C_1^3C_1^4C_1^5/C_3^{12} = 60/220$
 $P(1,2) = C_1^3C_2^4/C_3^{12} = 18/220$, $P(2,0) = C_2^3C_1^5/C_3^{12} = 15/220$
 $P(2,1) = C_2^3C_1^4/C_3^{12} = 12/220$, $P(3,0) = C_3^3/C_3^{12} = 1/220$

$P(i,j)$ 可表示为表 6-2 所示的矩阵表。

表 6-2 $P(i,j)$ 的矩阵表						
		变量 Y				
		0	1	2	3	$P_X(x)$
变 量 X	0	10/220	40/220	30/220	4/220	84/220
	1	30/220	60/220	18/220	0	108/220
	2	15/220	12/220	0	0	27/220
	3	1/220	0	0	0	1/220
$P_Y(y)$		56/220	112/220	48/220	4/220	1

定义 若 X, Y 是连续型随机变量, 则 X, Y 的联合概率密度函数 (joint probability density function), 记作 $f_{XY}(x,y)$, 或简记 $f(x,y)$, 可以图 6-5 (b) 表示, 满足下列条件:

- (1) $f(x,y) > 0$
- (2) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dx dy = 1$
- (3) $P(a \leqslant X \leqslant b,c \leqslant Y \leqslant d) = \int_a^b \int_c^d f(x,y) dy dx$

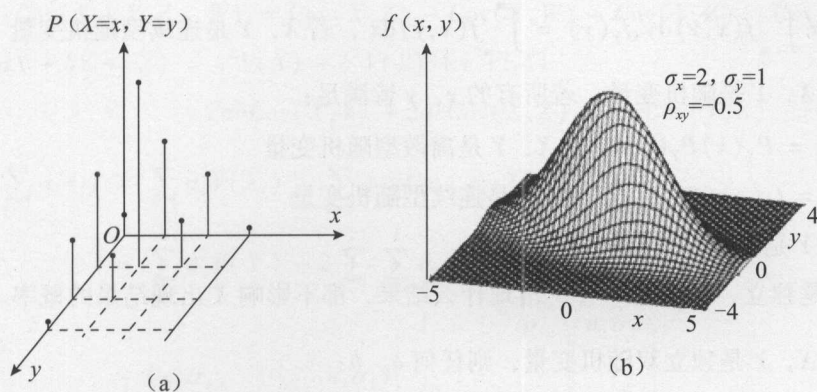


图 6-5 联合概率分布函数与联合概率密度函数

(a) 离散型联合概率分布函数; (b) 连续型联合概率密度函数

定义 若 X, Y 是双随机变量, 则 X, Y 的联合累积概率函数 (joint cumulative probability function), 记作 $F_{XY}(x, y)$, 或简记 $F(x, y)$, 如图 6-6 所示, 其公式为

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{i \leq x} \sum_{j \leq y} P(i, j) = \int_{s \leq x, t \leq y} f(s, t) dt ds$$

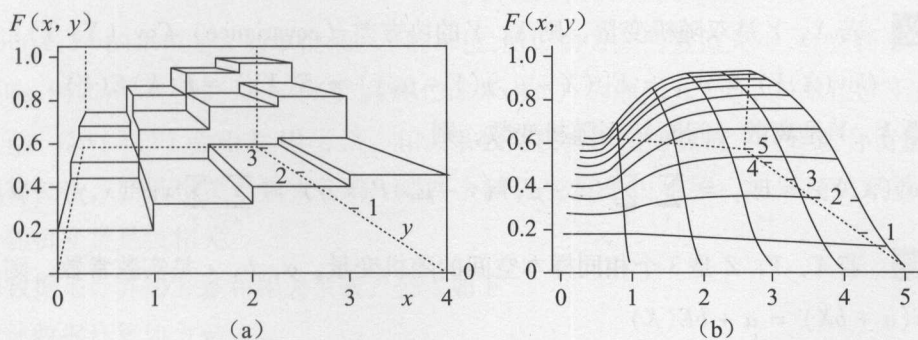


图 6-6 联合累积概率函数

(a) 离散型联合累积概率函数; (b) 连续型联合累积概率函数

性质 若 X, Y 是双随机变量, 则 X, Y 的联合累积概率函数有下列性质:

$$(1) \frac{\partial^2}{\partial x \partial y} F(x, y) = f(x, y)$$

$$(2) P(a \leq X \leq b, c \leq Y \leq d) = F(b, d) + F(a, c) - F(a, d) - F(b, c)$$

定义 若 X, Y 是随机变量, 则 X 的边际概率函数 (marginal probability function): 离散型 $P_X(x), P_Y(y)$ 或连续型 $f_X(x), f_Y(y)$:

$$P_X(x) = \sum_y P(x, y), P_Y(y) = \sum_x P(x, y), \text{ 若 } X, Y \text{ 是离散型随机变量}$$

$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy, f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$, 若 X, Y 是连续型随机变量

定义 X, Y 是随机变量, 若所有的 x, y 皆满足:

$P(x, y) = P_X(x)P_Y(y)$, 若 X, Y 是离散型随机变量

$f(x, y) = f_X(x)f_Y(y)$, 若 X, Y 是连续型随机变量

则 X 与 Y 是独立的 (independent)。

X 与 Y 是独立, 意思是不管 Y 出现什么结果, 都不影响 X 出现结果的概率。

定理 X, Y 是独立双随机变量, 则任何 a, b :

$$P(X \leq a, Y \leq b) = P\{X \leq a\}P\{Y \leq b\}, \quad F(a, b) = F_X(a)F_Y(b)$$

定义 若 X, Y 是双随机变量, $g(X, Y)$ 是 X, Y 的一个函数, 则 $g(X, Y)$ 的期望:

$$E[g(X, Y)] = \sum_x \sum_y g(x, y)P(x, y), \quad \text{若 } X, Y \text{ 是离散型随机变量}$$

$$E[g(X, Y)] = \iint_{x, y} g(x, y)f(x, y)dydx, \quad \text{若 } X, Y \text{ 是连续型随机变量}$$

定义 若 X, Y 是双随机变量, 则 X, Y 的协方差 (covariance) $\text{Cov}(X, Y)$:

$$\text{Cov}(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y)$$

如果 X, Y 是离散 (间断) 型随机变数, 则

$$\text{Cov}(X, Y) = \sigma_{XY} = \sum_x \sum_y (x - \mu_X)(y - \mu_Y)P(x, y) = \sum_x \sum_y xyP(x, y) - \mu_X\mu_Y$$

定理 若 X, Y, Z 是 3 个相同样本空间的随机变量, a, b, c 是实数常数, 则

$$1. E(a + bX) = a + bE(X)$$

$$2. V(a + bX) = b^2V(X)$$

$$3. E(aX + bY) = aE(X) + bE(Y)$$

$$4. V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab\text{Cov}(X, Y)$$

$$5. \text{Cov}(X, X) = V(X)$$

$$6. V(X \pm Y) = V(X) + V(Y) \pm 2\text{Cov}(X, Y)$$

$$7. \text{Cov}(X, Y) = \text{Cov}(Y, X)$$

$$8. \text{Cov}(a + bX, c + dY) = bd\text{Cov}(X, Y)$$

$$9. \text{Cov}(a, Y) = 0, \forall a \in R$$

$$10. \text{Cov}(X + Z, Y) = \text{Cov}(X, Y) + \text{Cov}(Z, Y)$$

$$11. \text{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{Cov}(X_i, Y_j)$$

例: $\text{Cov}(X + Y, Z - W) = \text{Cov}(X, Z) - \text{Cov}(X, W) + \text{Cov}(Y, Z) - \text{Cov}(Y, W)$

$$12. V(aX + bY + cZ) = a^2 V(X) + b^2 V(Y) + c^2 V(Z) \\ + 2ab \text{Cov}(X, Y) + 2ac \text{Cov}(X, Z) + 2bc \text{Cov}(Y, Z)$$

$$13. V\left(\sum_{i=1}^k a_i X_i\right) = \sum_{i=1}^k a_i^2 V(X_i) + \sum_{i \neq j} a_i a_j \text{Cov}(X_i, X_j) \\ = \sum_{i=1}^k a_i^2 V(X_i) + 2 \sum_{i=1}^k \sum_{j=i+1}^k a_i a_j \text{Cov}(X_i, X_j) \\ = (a_1 \sigma_1, \cdots, a_k \sigma_k) \begin{pmatrix} 1 & \cdots & \rho_{1k} \\ \vdots & \ddots & \vdots \\ \rho_{k1} & \cdots & 1 \end{pmatrix} \begin{pmatrix} a_1 \sigma_1 \\ \vdots \\ a_k \sigma_k \end{pmatrix}$$

14. 若 X, Y 是独立, 则

$$V(XY) = [E(X)]^2 V(Y) + [E(Y)]^2 V(X) + V(X) V(Y)$$

定义 若 X, Y 是双随机变量, 则 X, Y 的相关系数 (correlation coefficient):

$$\rho_{XY} = \text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$\text{Cov}(X, Y)$ 的单位是 X 的单位乘以 Y 的单位, 所以协方差的数值大小, 会受到 X 与 Y 的单位之影响。例如: $X =$ 身高, $Y =$ 体重, X 单位改米 (m) 为厘米 (cm), Y 单位改千克 (kg) 为克 (g), $\text{Cov}(X, Y)$ 会相差 10 万倍。相关系数是没有单位的数值, 比较适合衡量双随机变量的相关性。如果相关系数大于 0, 则两个随机变量是正相关; 如果相关系数小于 0, 则两个随机变量是负相关。

用数据来计算协方差和相关系数, 公式如下

总体数据计算协方差 σ_{XY}

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)$$

样本数据计算协方差 S_{XY}

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \left(\sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \right)$$

样本数据计算相关系数 r

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum (x_i - \bar{x})^2)(\sum (y_i - \bar{y})^2)}}$$

定理 若 X, Y 是双随机变量, 则 X, Y 的相关系数 ρ_{XY} :

$$1. -1 \leq \rho_{XY} \leq +1$$

2. $|\rho_{XY}| = 1$ 若且唯若存在 a 和 $b \neq 0$, 使 $P(Y = a + bX) = 1$
若 $\rho_{XY} = 1$, 则 $b > 0$; 若 $\rho_{XY} = -1$, 则 $b < 0$ 。
 ρ_{XY} 只是衡量随机变量 X, Y 的“线性”关系。

定理 若 X, Y 是独立的双随机变量, 则 $Cov(X, Y) = 0, \rho_{XY} = 0$

以上逆定理并不成立。即若 $Cov(X, Y) = 0$, 但 X, Y 不一定是独立的。因为 $\rho_{XY} = 0$, 只是表示 X 与 Y 完全没有“线性”相关, 并非完全没有相关。例如: X 与 Z 是独立的连续均匀分布定义在 $[-1, +1]$, $E(X) = E(Z) = 0, E(X^3) = 0$, 令 $Y = X^2 + Z$, 所以 X, Y 不是独立的, 但是:

$Cov(X, Y) = E(XY) - E(X)E(Y) = E[X(X^2 + Z)] = E(X^3) + E(X)E(Z) = 0$
相关系数的推论统计请见第 13 章相关分析。

随机变量相关的分类如图 6-7 所示。

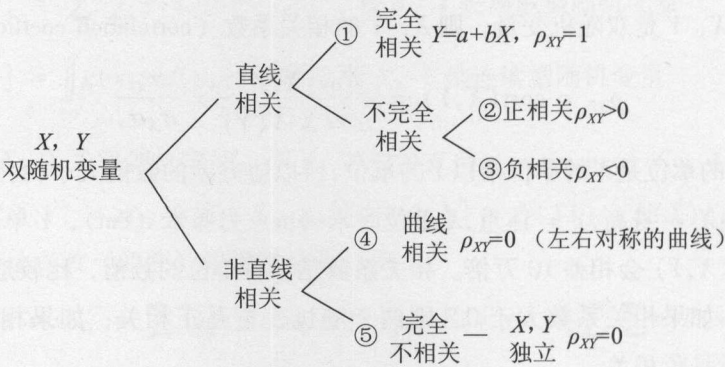


图 6-7 随机变量相关的分类

例题 6.7 随机变量 X, Y 的联合概率密度函数如表 6-3 所示。

表 6-3 X, Y 的联合概率密度函数

		变量 Y				$P_X(x)$
		1	2	3	4	
变量 X	1	0.12	0.03	0.06	0.09	0.30
	2	0.20	0.05	0.10	0.15	0.50
	3	0.08	0.02	0.04	0.06	0.20
	$P_Y(y)$	0.40	0.10	0.20	0.30	1

证明随机变量 X, Y 是独立的。

解答: 随机变量 X, Y 是独立的证明如表 6-4 所示。

表 6-4 X, Y 是独立的证明

变量 X	变量 Y				$P_X(x)$
	1	2	3	4	
	(0.3) (0.4) =0.12	(0.3) (0.1) =0.03	(0.3) (0.2) =0.06	(0.3) (0.3) =0.09	0.30
	(0.5) (0.4) =0.20	(0.5) (0.1) =0.05	(0.5) (0.2) =0.10	(0.5) (0.3) =0.15	0.50
	(0.2) (0.4) =0.08	(0.2) (0.1) =0.02	(0.2) (0.2) =0.04	(0.2) (0.3) =0.06	0.20
$P_Y(y)$	0.40	0.10	0.20	0.30	1

所以随机变量 X, Y 是独立的。

如果随机变量 X, Y 是独立的，则变量 X 与变量 Y 的联合概率分布的矩阵表（列联表），每列是成比例的，每行也是成比例的，即

$(0.12:0.2:0.08) = (0.03:0.05:0.02) = (0.06:0.1:0.04)$

事件之独立只是某行和某列是独立的，例如： $\{X=2\}$ 和 $\{Y=3\}$ 是独立的。

例题 6.8—6.12 （见网络资源）

例题 6.13 投资组合

假设你有 100 万资金，可以投资下列 A、B、C 3 种股票，各种股票的报酬率是一个随机变量，已知这 3 种股票的期望报酬（期望值）、风险（标准差）及相关系数如表 6-5 所示。

表 6-5 3 种股票的期望报酬、风险及相关系数

股票	随机变量	期望报酬	标准差	相关系数		
		$E(X) / (\% / \text{年})$	风险 $\sigma / \%$	A	B	C
A	X	9.1	16.5		-0.22	0.13
B	Y	12.1	15.8	-0.22		0.41
C	Z	11.2	13.9	0.13	0.41	

请问投资的组合应该如何分配？

解答：一般理性或风险避免者会选择期望报酬大，风险小的股票。我们注意股票 A，其期望报酬最小，风险最大。本来不应该投资股票 A，但是股票 A 和股票 B 的相关系数为负数（例如国内航空公司和高铁公司，获利的相关系数为负数）。

所以我们的投资组合 W 分配到股票 A 和股票 B：

$$W = fX + (1 - f)Y, \quad 0 \leq f \leq 1$$

$$E(W) = 9.1f + 12.1(1 - f)$$

$$V(W) = 16.5^2 f^2 + 15.8^2 (1 - f)^2 + 2f(1 - f)(16.5)(15.8)(-0.22)$$

$$\sigma_w = \sqrt{272.25 f^2 + 249.64 (1 - f)^2 + 2f(1 - f)(16.5)(15.8)(-0.22)}$$

$f, E(W), \sigma(W)$ 的关系如下表所示。

表 6-6 $f, E(W), \sigma_w$ 的关系

f	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$E(W)$	12.1	11.8	11.5	11.2	10.9	10.6	10.3	10.0	9.7	9.4	9.1
σ_w	15.8	13.95	12.34	11.08	10.29	10.08	10.51	11.47	12.88	14.58	16.5

将投资组合 W 的期望报酬和风险画出如图 6-8，我们注意到投资组合 $f=0.2$ ：20% 股票 A 和 80% 股票 B，期望报酬和风险都比股票 C 好。

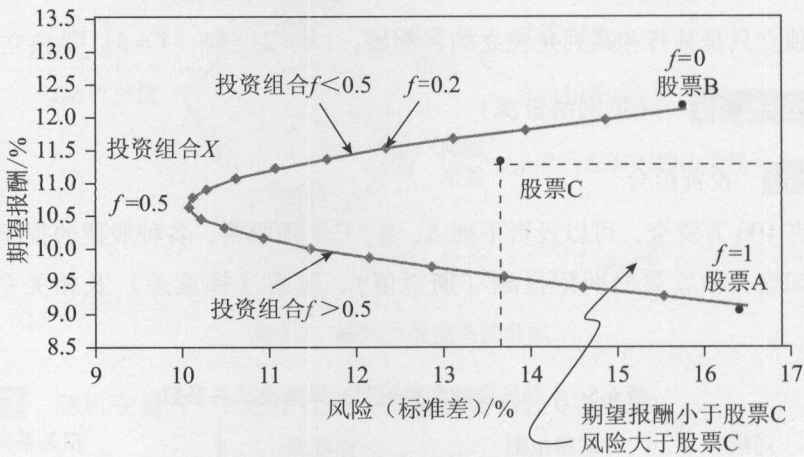


图 6-8 投资组合的期望报酬和标准差

当 $0 \leq f \leq 0.5$ ，投资组合 $W = fX + (1 - f)Y$ 都是可以考虑的。
如果考虑股票 A 和股票 C，作为投资组合，是否有更大的期望报酬和更小的风险？
投资组合的期望值与方差：

(1) 若 $W = w_1X_1 + w_2X_2$ 是股票 X_1, X_2 的投资组合， w_1, w_2 是投资比重，则
期望值 $E(W) = w_1E(X_1) + w_2E(X_2)$

方差 $V(W) = w_1^2V(X_1) + w_2^2V(X_2) + 2w_1w_2Cov(X_1, X_2)$

(2) 若 $W = \sum_{i=1}^k w_iX_i$ 是股票 X_1, X_2, \dots, X_k 的投资组合， w_1, w_2, \dots, w_k 是投资比重，则

$$\text{期望值 } E(W) = \sum_{i=1}^k w_i E(X_i)$$

$$\text{方差 } V(W) = \sum_{i=1}^k w_i^2 V(X_i) + 2 \sum_{i=1}^k \sum_{j=i+1}^k w_i w_j \text{Cov}(X_i, X_j)$$

6.5 中文统计应用

6.5.1 一个离散型随机变量的统计值（例题 6.4）

执行“一个离散型随机变量的统计值”的操作示意图和结果分别如图 6-9 和图 6-10 所示。

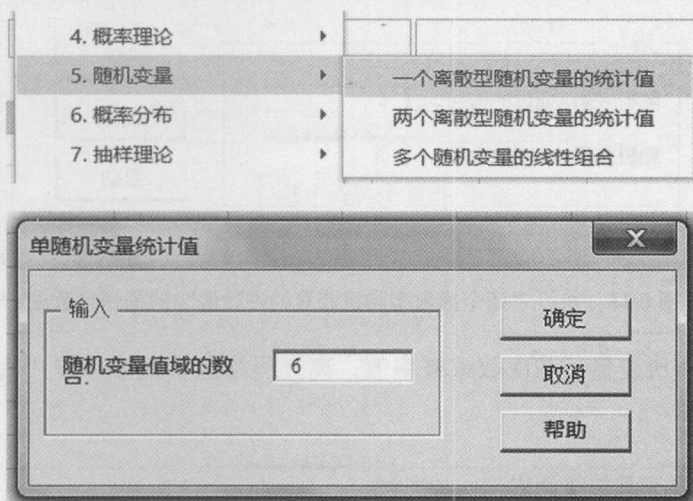


图 6-9 执行“一个离散型随机变量的统计值”的操作示意图

随机变量的统计值					
#X=	6				
输入：	随机变量值	概率值		随机变量 X 的统计值	
	x	P(x)	F(x)		
	1.00	0.17	0.166667	期望值(均值)	3.5
	2.00	0.17	0.333333	方差	2.916667
	3.00	0.17	0.5	标准误差	1.707825
	4.00	0.17	0.666667	偏态	0
	5.00	0.17	0.833333	峰态	1.731429
	6.00	0.17	1		
		1/6		重新计算	

图 6-10 执行“一个离散型随机变量的统计值”的结果

6.5.2 两个离散型随机变量的统计值（例题 6.12）

执行“两个离散型随机变量的统计值”的操作示意图如图 6-11 所示。

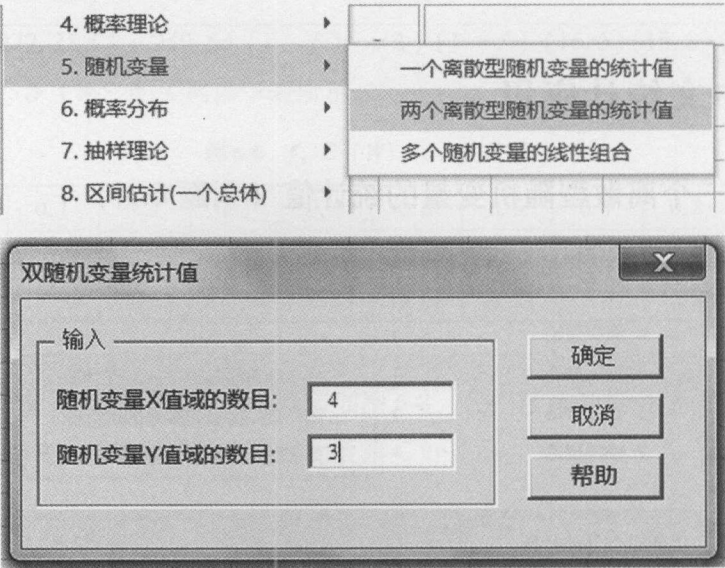


图 6-11 执行“两个离散型随机变量的统计值”的操作示意图

输入 X 与 Y 随机变量值与次数或概率值，离开浅绿色单元格，按“重新计算”，如图 6-12 所示。

双随机变量的统计值		(#X,#Y)=		4	3	
输入数据:		浅蓝色储存格输入数据，浅黄色储存格是计算结果				
		Y (输入Y 随机变量值)				
可输入次数或概率		30.00	40.00	50.00	边际总和	重新计算
输入 X 值	0.00	0.01	0.02	0.05	0.08	
	1.00	0.03	0.06	0.10	0.19	
	2.00	0.18	0.21	0.15	0.54	
	3.00	0.07	0.08	0.04	0.19	
	边际总和	0.29	0.37	0.34	1.00	
E(X)		1.84		V(X)	0.6744	
E(Y)		40.5		V(Y)	62.75	
E(XY)		72.8				
Cov(X,Y)		-1.72				
ρXY		-0.2644				

图 6-12 执行“两个离散型随机变量的统计值”的结果

6.6 本章流程图

本章流程图如图 6-13 所示。

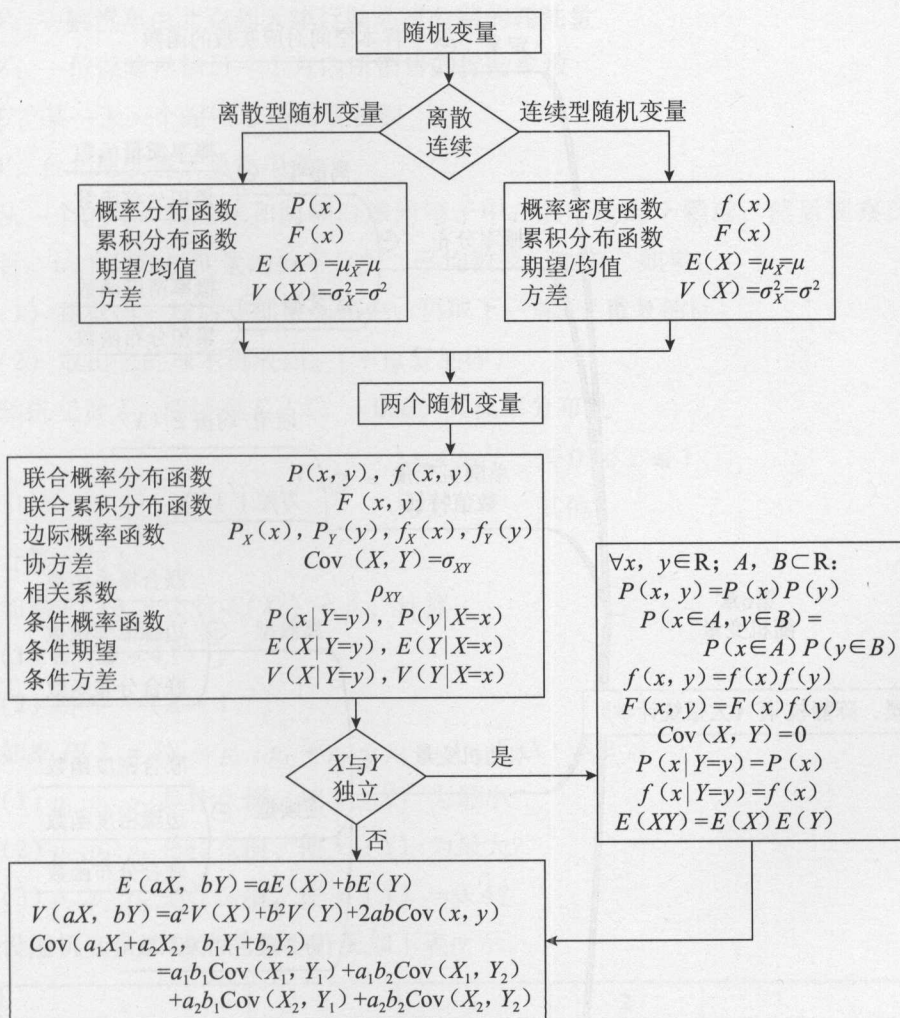


图 6-13 第 6 章的流程图

6.7 本章思维导图

本章思维导图如图 6-14 所示。

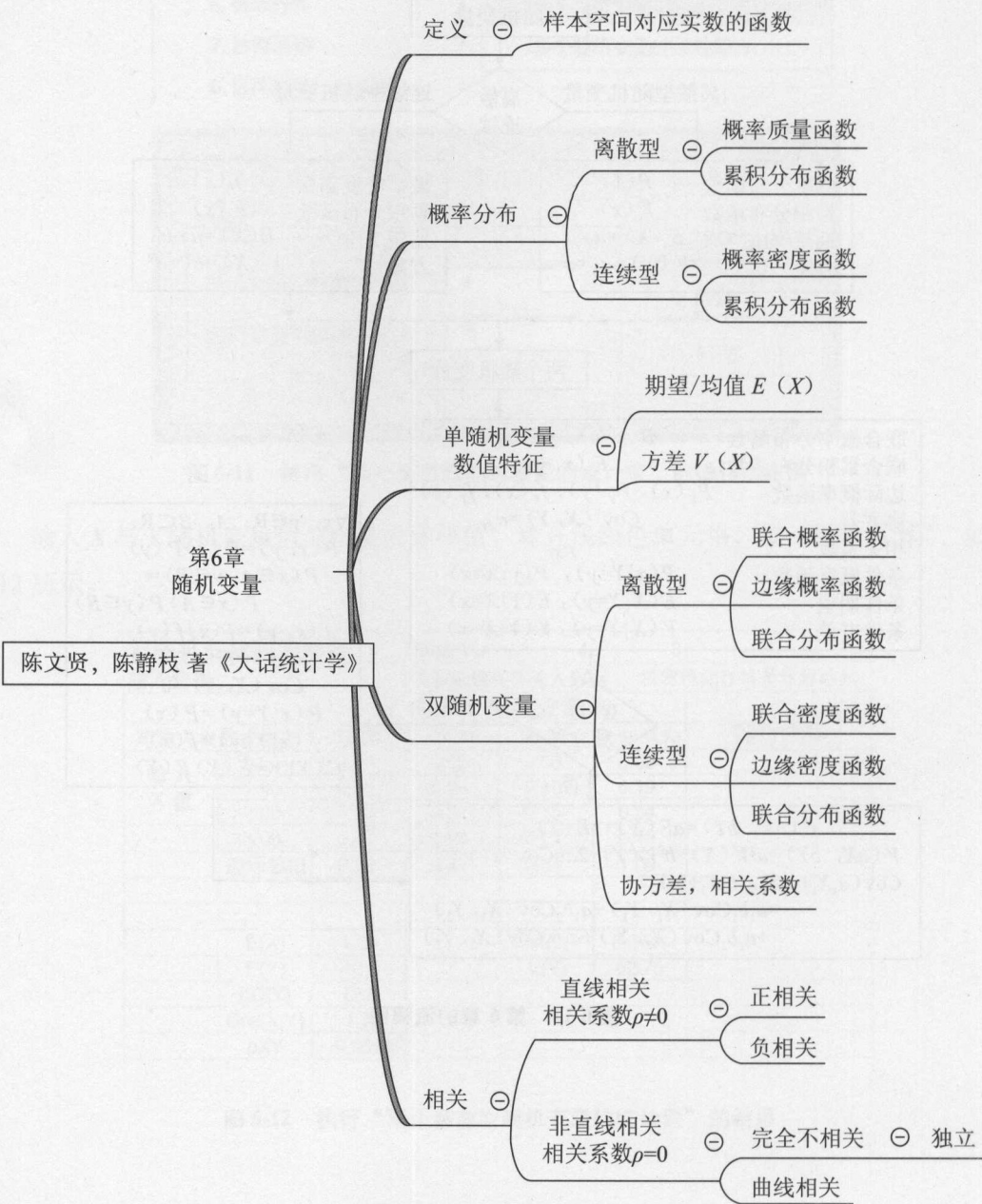


图 6-14 第 6 章思维导图

习 题

1. 下列变量中, 那些是离散型随机变量? 那些是连续型随机变量?

X : 一批产品所含不良产品数

Y : 一辆汽车由北京到天津行驶高速公路的耗油量

Z : 一位保险推销员一个月内所销售的保险单数

U : 某一天一个超级市场的营业额

V : 台北地区每天总耗电量

2. 从一个包含4颗红球和两颗白球的箱子中, 顺序取出3颗球, 然后观察白球的数目。试求此随机变量的概率分配、平均数及标准差, 如果:

(1) 每取出一球后立即放回箱中, 再取下一球。(重复抽样)

(2) 取出来的球不再放回。(不重复抽样)

3. 随机变量 X , 期望值 $E(X) = 0.6$, 其概率分布是

$$f(x) = \begin{cases} a + bx^2, & \text{若 } 0 \leq x \leq 1 \\ 0, & \text{其他} \end{cases}$$

计算 a 与 b 。

4. 如果 $E(X) = 2$ 且 $E(X^2) = 8$, 计算:

(1) $E[(2 + 4X)^2]$

(2) $E[X^2 + (X + 1)^2]$

5. 如果 $P(X = i) = p_i$, $p_1 + p_2 + p_3 = 1$, $E(X) = 2$ 。

(1) p_1, p_2, p_3 是什么值, 使 $V(X)$ 为最小?

(2) p_1, p_2, p_3 是什么值, 使 $V(X)$ 为最大?

(3) p_1, p_2, p_3 是什么值, 使 $V(X) = 0.4$?

6. 设随机变量 X 的离散概率分配如下表所示。

x	0	1	2	3	4
概率 $P(x)$	0.2	0.4	0.3	0.05	0.05

(1) 计算 $P(X > 1)$, $P(1 < X < 3)$ 。

(2) 计算 $E(X)$, $V(X)$ 。

(3) 设 $Y = (X - 3)(X - 3)$, 求 Y 的概率分布。

(4) 由 Y 的概率分布, 求 $E(Y)$ 与 $V(Y)$ 。

其他习题请下载。



第7章

概率分布

故兵无常势，水无常形，能因敌变化而取胜，谓之神。

故五行无常胜，四时无常位，日有短长，月有死生。

——《孙子兵法·虚实篇》

人法地，地法天，天法道，道法自然。

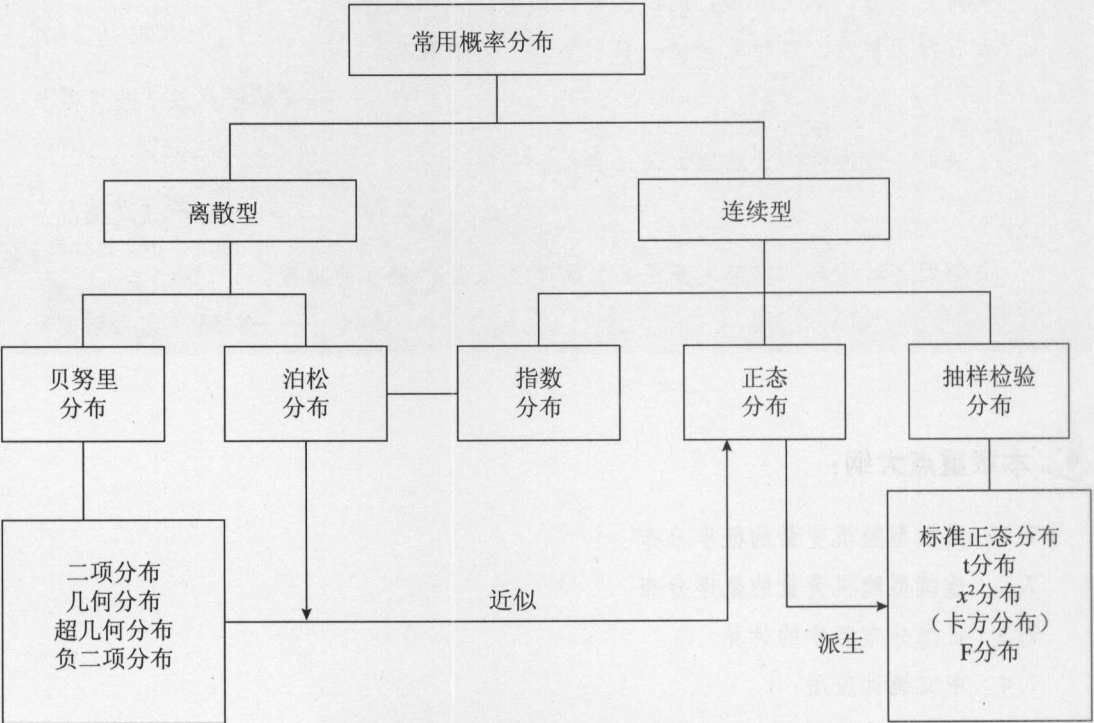
——老子《道德经》

夫心术之动远矣，文情之变深矣，源奥而派生，根盛而颖峻。

——刘勰《文心雕龙》

本章重点大纲：

- 7.1 离散型随机变量的概率分布
- 7.2 连续型随机变量的概率分布
- 7.3 正态分布概率的计算
- 7.4 中文统计应用
- 7.5 本章思维导图
- 7.6 本章流程图



本章概念图（详细关联请见本章流程图）

7.1 离散型随机变量的概率分布

在上一章我们提到：每个随机变量会有对应的概率函数、期望值、方差、偏度系数、峰度系数等。期望值、方差、偏度系数、峰度系数等，我们称为特征值。

随机变量 X 对应的概率分布函数 F ，称为随机变量服从该概率分布，记作 $X \sim F$ 。

除了任意离散型概率分布，每个概率分布函数有一个或一个以上的“参数”，决定这个概率分布函数的这些特征值。参数是确定值，在本章及下一章是已知值，在第9章以后是未知值。

以下我们要介绍的离散型随机变量的概率分布函数有：任意离散型概率分布，离散均匀概率分布 (discrete uniform distribution)，贝努里分布 (Bernoulli distribution)，二项分布 (binomial distribution)，泊松分布 (Poisson distribution)，超几何分布 (hyper geometry distribution)，几何分布 (geometry distribution)，负二项分布 (negative Binomial distribution)。

7.1.1 任意离散型概率分布

任意离散型概率分布，通常是用表格或条形图表示概率分布函数，例如：

表 7-1 概率分布表

实数值 x	概率值 $P(x)$
1	0.2
2	0.3
3	0.4
4	0.1

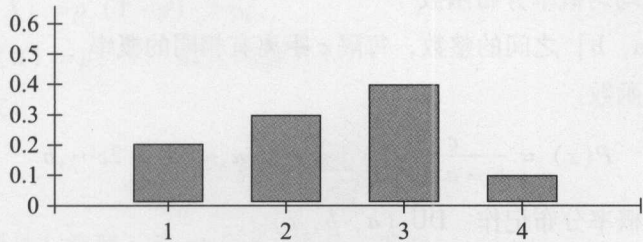


图 7-1 概率分布条形图

任意离散型概率分布没有直接的公式，表达其均值与方差等特征值。只有代入定义。

7.1.2 离散均匀概率分布

离散均匀分布 (discrete uniform distribution) 是随机变量每个实数, 如果概率不等于 0, 则有相同的概率。

1. 特殊类型的均匀概率分布函数

概率分布在 1 到 N 之间的整数, 有相同的概率, 如图 7-2 所示。

(1) 概率分布函数

$$P(x) = \frac{1}{N}, x = 1, 2, \dots, N$$

(2) 参数 = N (正整数)。

(3) 均值

$$E(X) = \frac{N+1}{2}$$

(4) 方差: $V(X) = \frac{N^2 - 1}{12}$ 。

(5) 偏度系数: $S(X) = 0$ (对称型)。

(6) 峰度系数: $K(X) = \frac{3}{5} \left(3 - \frac{4}{N^2 - 1} \right) < \frac{9}{5}$ (平峰型)。

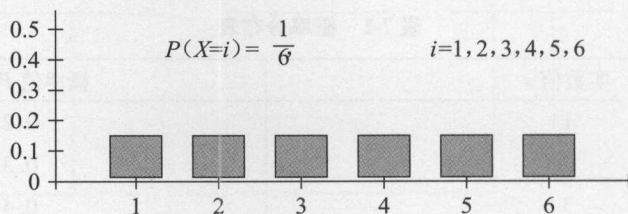


图 7-2 特殊类型的均匀概率分布函数

2. 一般类型的均匀概率分布函数

概率分布在 $[a, b]$ 之间的整数, 每隔 c 距离有相同的概率。

(1) 概率分布函数

$$P(x) = \frac{c}{b - a + c}, \quad x = a, a + c, a + 2c, \dots, b$$

(2) 离散均匀概率分布记作 $DU(a, b, c)$ 。

(3) 参数 = a, b, c (都是正整数), $b = kc, k = [(b - a + c)/c] - 1$ 。

(4) 均值

$$E(X) = \frac{a+b}{2}$$

(5) 方差

$$V(X) = \frac{(b - a + c)^2 - c^2}{12}$$

7.1.3 贝努里分布

贝努里分布 (Bernoulli distribution) 或称两点分布, $0 \sim 1$ 分布的随机变数: 一次“成功或失败”试验, 失败的变数值为 0, 成功的变数值为 1, 成功率为 p (以后在估计或检验中称为“总体比例”)。贝努里分布如图 7-3 所示。

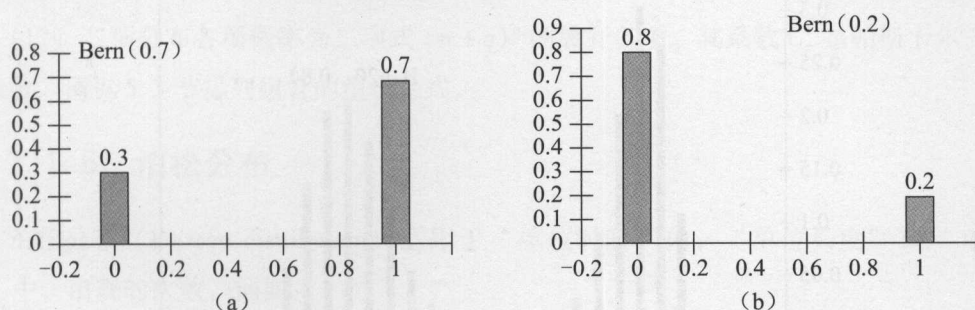


图 7-3 贝努里分布

(a) Bern (0.7); (b) Bern (0.2)

(1) 概率分布函数

$$P(x) = \begin{cases} 1 - p, & \text{若 } x = 0 \\ p, & \text{若 } x = 1 \\ 0, & \text{其他} \end{cases}$$

(2) 贝努里分布记作 Bern (p)

(3) 参数 = p , $0 \leq p \leq 1$, 通常令 $q = 1 - p$ 。

(4) 均值: $E(X) = p$ 。

(5) 方差: $V(X) = p(1 - p) = pq$ 。

(6) 众数: $M(X) = p$, 取其四舍五入。

(7) 偏度系数

$$S(X) = \frac{1 - 2p}{\sqrt{p(1 - p)}}$$

若 $p < 0.5$, 则是右偏型; 若 $p > 0.5$, 则是左偏型。

(8) 峰度系数

$$K(X) = 3 + \frac{1 - 6p + 6p^2}{p(1 - p)}$$

若 $p=0.5$, 则 $K(X)=1$ 为平峰型。若 p 越接近 0 或 1, 则越变成为尖峰型。

7.1.4 二项分布

二项分布 (binomial distribution) 是 n 个独立贝努里分布 $\text{Bern}(p)$ 之和。随机变数 X 是: n 次独立的“成功或失败”试验中, 成功的次数。若 $Y_i \sim \text{Bern}(p)$, 且 Y_i 是独立的, 则 $X = \sum Y_i \sim B(n, p)$ 。二项分布如图 7-4 所示。

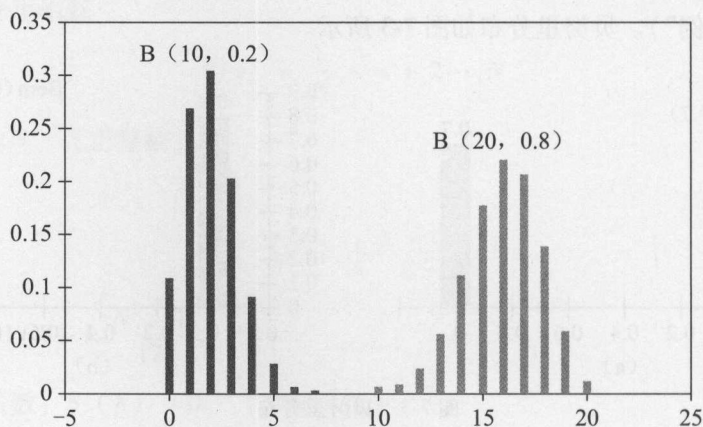


图 7-4 二项分布

(1) 概率分布函数

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

(2) 二项分布记作 $B(n, p)$ 。

(3) 参数 = $n, p, 0 \leq p \leq 1$, n 为大于 1 的正整数, 通常令 $q = 1 - p$ 。

(4) 均值: $E(X) = np$ 。

(5) 方差: $V(X) = np(1-p) = npq$ 。

(6) 众数: $M(X) = np$, 取其四舍五入。

(7) 偏度系数

$$S(X) = \frac{1-2p}{\sqrt{np(1-p)}} = \frac{1-2p}{\sqrt{npq}}$$

若 $p < 0.5$, 则是右偏型; 若 $p > 0.5$, 则是左偏型。

若 p 越接近 0.5, 则越对称; 若 p 越接近 0 或 1, 则右偏或左偏越明显。

(8) 峰度系数

$$K(X) = 3 + \frac{1-6p+6p^2}{np(1-p)} = 3 + \frac{1-6pq}{npq}$$

若 $p=0.5$, 则为平峰型。若 $pq > 1/6$, 即 p 越接近 0.5, 则越成为平峰型。若 $pq < 1/6$, 即若 p 越接近 0 或 1, 则越成为尖峰型。但是 n 越大, 则越接近正态峰。

(9) 相同参数 p 的独立二项分布是可加性, 即若 $X \sim B(n, p)$, $Y \sim B(m, p)$, 且 X 和 Y 为独立的, 则 $X+Y \sim B(n+m, p)$ 。

(10) 当二项分布 $n \rightarrow \infty, p \rightarrow 0$ 且 np 为常数, 则可以利用泊松分布求其近似值, 泊松分布参数: $\lambda = np$ 。即 $B(n, p) \rightarrow \text{Pois}(np)$ 。

(11) 若二项分布 $n \rightarrow \infty$ 且 $np > 5, n(1-p) > 5$, 则可以利用正态分布求其近似值, 正态分布参数: $\mu = np, \sigma^2 = np(1-p) = npq$ 。即 $B(n, p) \rightarrow N(np, npq)$ 。

(12) 二项分布各项概率为二项式 $(p+q)^n$ 的展开各项。其系数 C_x^n 是帕斯卡尔三角形的系数。请见 5.3 节排列组合的组合公式。

7.1.5 泊松分布

泊松分布 (Poisson distribution) 适用于“单位时间”内, “单位长度”或“单位面积”中, 出现的次数, 例如:

(1) 一年内某医院 (或城市), 癌症死亡人数。

(2) 一周内工厂, 机器故障次数。

(3) 一天内某超市, 进入顾客的人数。

(4) 一小时内, 打进电话的次数。

(5) 一页书内, 错字的数目。

(6) 森林中一平方千米内, 动物的数目。

(7) 每 100 米的街上, 餐厅的数目。

泊松分布如图 7-5 所示。

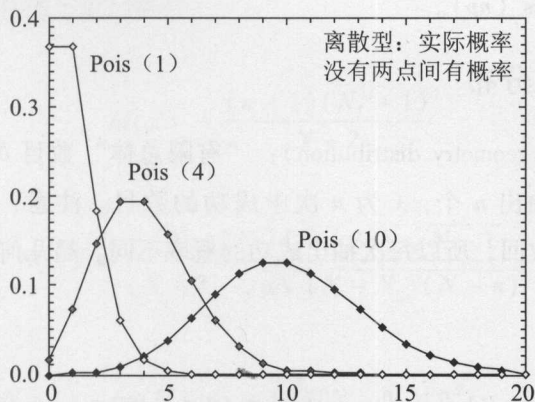


图 7-5 泊松分布

(1) 概率分布函数

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, 2, 3, \dots, \infty$$

(2) 泊松分布记作 $\text{Pois}(\lambda)$ 。(3) 参数 $= \lambda, \lambda > 0$ 。(4) 均值: $E(X) = \lambda$ 。(5) 方差: $V(X) = \lambda$ 。(6) 众数 $M(X) = \lambda$ 附近的整数。

(7) 偏度系数

$$S(X) = \frac{1}{\sqrt{\lambda}} > 0 \quad (\text{右偏型})$$

(8) 峰度系数

$$K(X) = 3 + \frac{1}{\lambda} \quad (\text{尖峰型})$$

(9) 独立的泊松分布是线性可加性, 即若 $X \sim \text{Pois}(\lambda_1), Y \sim \text{Pois}(\lambda_2)$, 且 X 和 Y 为独立的, 则 $aX + bY \sim \text{Pois}(a\lambda_1 + b\lambda_2)$

(10) 若单位时间 (例如一小时) 内到达的人数是泊松分布, 则相邻到达者的间隔时间 (两个人到达的距离时间) 是指数分布。

(11) 若单位时间 (例如一小时) 内到达的人数 X 是泊松分布 $X \sim \text{Pois}(\lambda)$, 则任何 t 时间 (例如 3h) 内到达的人数 Y 是泊松分布 $Y \sim \text{Pois}(\lambda t)$ [例如 $X \sim \text{Pois}(3\lambda)$]。

(12) 若泊松分布 $\lambda > 5$, 则可以利用正态分布求其近似值, 正态分布的参数: $\mu = \lambda, \sigma^2 = \lambda$ 。即 $\text{Pois}(\lambda) \rightarrow N(\lambda, \lambda)$ 。

(13) 若二项分布 $n \rightarrow \infty, p \rightarrow 0$, 则可以利用泊松分布求其近似值, 泊松分布参数: $\lambda = np$ 。即 $B(n, p) \rightarrow \text{Pois}(np)$ 。

7.1.6 超几何分布

超几何分布 (hypergeometry distribution): “有限总体”数目 N , 其中成功的个数为 N_1 , 以“不返回式”抽出 n 个, x 为 n 次中成功的数目。注意: 第一次试验成功率为 N_1/N , 因为抽出后不放回, 所以每次抽出成功的概率不同。超几何分布如图 7-6 所示。

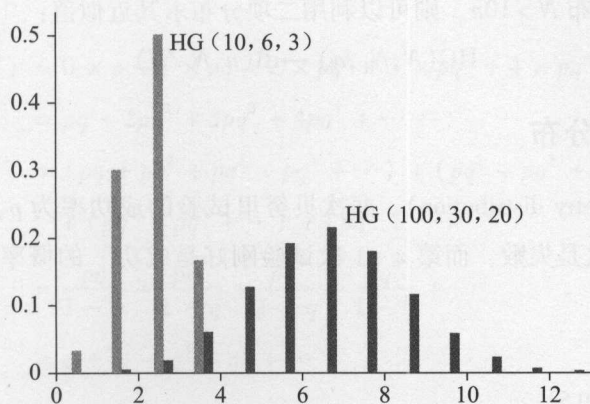


图 7-6 超几何分布

(1) 概率分布函数

$$P(x) = \frac{\binom{N_1}{x} \binom{N - N_1}{n - x}}{\binom{N}{n}}, x = \max\{n - N + N_1, 0\}, \dots, \min\{N_1, n\}$$

(2) 超几何分布记作 $HG(N, N_1, n)$ 。

(3) 参数 = N, N_1, n (都是正整数)。

(4) 均值

$$E(X) = \frac{nN_1}{N}$$

(5) 方差

$$V(X) = n \left(\frac{N_1}{N} \right) \left(1 - \frac{N_1}{N} \right) \left(\frac{N - n}{N - 1} \right)$$

(对照第 8 章有限总体不返回抽样)

(6) 众数

$$M(X) = \frac{(n + 1)(N_1 + 1)}{N + 2}$$

(7) 偏度系数

$$S(X) = \frac{(N - 2N_1)(N - 2n) \sqrt{N - 1}}{(N - 2) \sqrt{nN_1(N - N_1)(N - n)}}$$

(8) 峰度系数

$$K(X) = \frac{N^2(N - 1) \{ N(N + 1) - 6n(N - n) + 3 \frac{N_1}{N^2} (N - N_1) [N^2(n - 2) - Nn^2 + 6n(N - n)] \}}{(N - 2)(N - 3)nN_1(N - N_1)(N - n)}$$

(9) 若超几何分布 $N > 10n$ ，则可以利用二项分布求其近似值：

$$HG(N, N_1, n) \rightarrow B(n, N_1/N)$$

7.1.7 几何分布

几何分布 (geometry distribution)：每次贝努里试验的成功率为 p ， $P(x)$ 为 $x+1$ 次独立试验，“前面 x 次是失败，而第 $x+1$ 次试验刚好是成功”的概率。几何分布如图 7-7 所示。

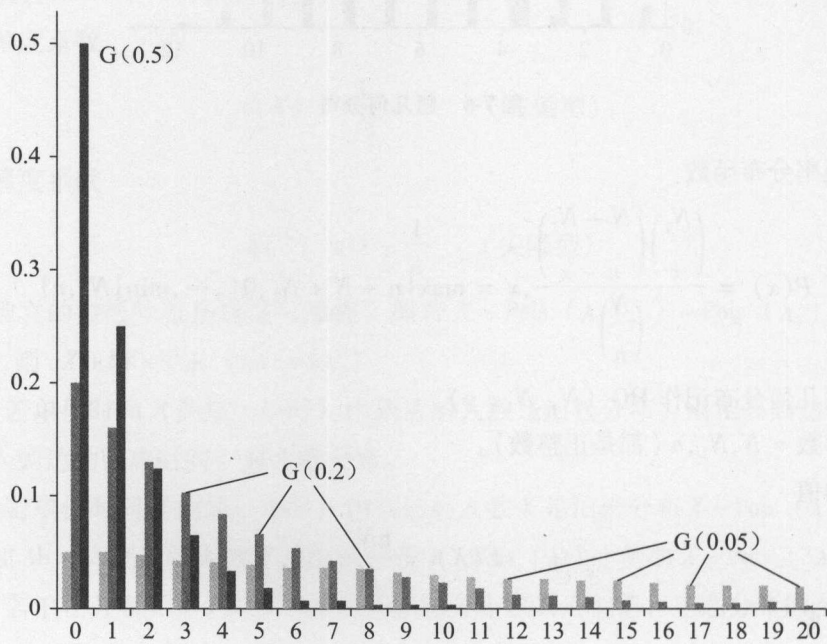


图 7-7 几何分布

(1) 概率分布函数

$$P(x) = p(1 - p)^x, x = 0, 1, 2, 3, \dots, \infty$$

x	0	1	2	3	4	5	...
$P(x)$	p	pq	pq^2	pq^3	pq^4	pq^5	...

(2) 几何分布记作 $G(p)$

(3) 参数 $= p, 0 \leq p \leq 1$ ，令 $q = 1 - p$ 。

(4) 均值

$$E(X) = \frac{q}{p}$$

证明:

$$\begin{aligned}
 E(X) &= 0 \times p + 1 \times pq + 2 \times pq^2 + 3 \times pq^3 + 4 \times pq^4 + \cdots \\
 &= pq + 2pq^2 + 3pq^3 + 4pq^4 + \cdots \\
 &= (pq + pq^2 + pq^3 + pq^4 + \cdots) + (pq^2 + pq^3 + pq^4 + \cdots) \\
 &\quad + (pq^3 + pq^4 + \cdots) + (pq^4 + \cdots) \\
 &= \frac{pq}{1-q} + \frac{pq^2}{1-q} + \frac{pq^3}{1-q} + \frac{pq^4}{1-q} + \cdots \\
 &= q + q^2 + q^3 + q^4 + \cdots \\
 &= \frac{q}{1-q} \\
 &= \frac{q}{p}
 \end{aligned}$$

(5) 方差

$$V(X) = \frac{q}{p^2}$$

(6) 偏度系数

$$S(X) = \frac{1+q}{\sqrt{q}} > 0 \quad (\text{右偏型})$$

(7) 峰度系数

$$K(X) = 3 + \frac{p^2 + 6q}{q} \quad (\text{尖峰型})$$

(8) 几何分布的各项概率值是几何级数: $p, pq, pq^2, pq^3, \cdots$, 故称几何分布。

(9) 独立的几何分布相加以后是负二项分布, 即

若 $X_i \sim G(p), i = 1, \cdots, n$, 且 X_i 为独立, 则 $Y = \sum_{i=1}^n X_i \sim \text{NB}(n, p)$

(10) 几何分布在离散型概率分布中唯一具有“无记忆性质”(memoryless property)。无记忆性质的意义: 不管过去 t 时间内事件是否发生, 从现在开始 s 时间内发生事件的概率和过去无关。例如: 不管过去 2h 有 100 个客人也好, 有 500 个客人也好, 从现在起 1h 内有 50 个客人的概率, 都是一样。无记忆性的数学定义:

$P(X > s+t | X > t) = P(X > s)$, 对所有的 $s = 0, 1, 2, \cdots, t = 0, 1, 2, \cdots$, 均成立。

(11) 几何分布有的定义: 每次贝努里试验的成功率为 p , $P(x)$ 为 x 次独立试验, “前面 $x-1$ 次是失败, 而第 x 次试验刚好是成功” 的概率。概率分布函数

$$P(x) = p(1-p)^{x-1}, x = 1, 2, \cdots, \infty$$

x	1	2	3	4	5	6	...
$P(x)$	p	pq	pq^2	pq^3	pq^4	pq^5	...

均值 $E(X) = 1/p$ ，方差 $V(X) = q/p^2$ 。

例题 7.1 圣彼得堡悖论

圣彼得堡悖论 (St. Petersburg paradox) 是决策论中的一个矛盾的悖论，1713 年由尼古拉·贝努里 (Nicolaus I Bernoulli) 提出。1738 年，其堂弟丹尼尔·贝努里 (Daniel Bernoulli) 在圣彼得堡学院论文集，以效用理论来解答这个问题。他们都是提出贝努里分布的雅各布·贝努里 (Jakob I. Bernoulli) 的侄子。

圣彼得堡悖论是一个赌局：掷一个公正的硬币，若第一次掷出正面，你就赚 2 元，赌局结束。若第一次掷出反面，那就要再掷一次，若第二次掷的是正面，你赚 4 元，赌局结束。若第二次掷出反面，那就要掷第三次，若第三次掷的是正面，你赚 8 元，赌局结束。若第四次掷的是正面，你赚 16 元，如此类推。问题是，你最多肯付多少钱参加这个赌局？

解答：概率分布函数

x	2	4	8	16	32	64	...
$P(x)$	1/2	1/4	1/8	1/16	1/32	1/64	...

$$\begin{aligned} \text{期望 } E(X) &= 2 \times \frac{1}{2} + 4 \times \frac{1}{4} + 8 \times \frac{1}{8} + 16 \times \frac{1}{16} + 32 \times \frac{1}{32} + \dots \\ &= 1 + 1 + 1 + 1 + 1 + \dots = \infty \end{aligned}$$

期望是无限大，表示你可以几千亿的钱来赌这个赌局。

方差 $V(X) = \sum (x - \infty)^2 P(x) = \sum \infty^2 P(x) = \infty$

通常做决策，是选择期望值最大，方差最小。圣彼得堡悖论是期望值是无限大，方差也是无限大，这就是矛盾的地方。另外，就是用期望效用理论来解释。

7.1.8 负二项分布

负二项分布 (negative binomial distribution)：每次试验的成功率为 p ，到成功 k 次时即终止的独立贝努里试验中，失败次数 x 的分布。 $P(x)$ 为 $x+k$ 次独立试验，“有 k 次成功， x 次失败，而第 $x+k$ 次试验刚好是成功”的概率。例如，连续掷一个硬币，在第 3 次 ($k=3$) 出现前，出现 x 次反面的次数之概率。负二项分布如图 7-8 所示。

(1) 概率分布函数

$$P(x) = \binom{x+k-1}{x} p^k (1-p)^x, x = 0, 1, 2, \dots$$

(2) 负二项分布记作 $NB(k, p)$ 。

(3) 参数 $= p, q = 1 - p$, k 为正整数。

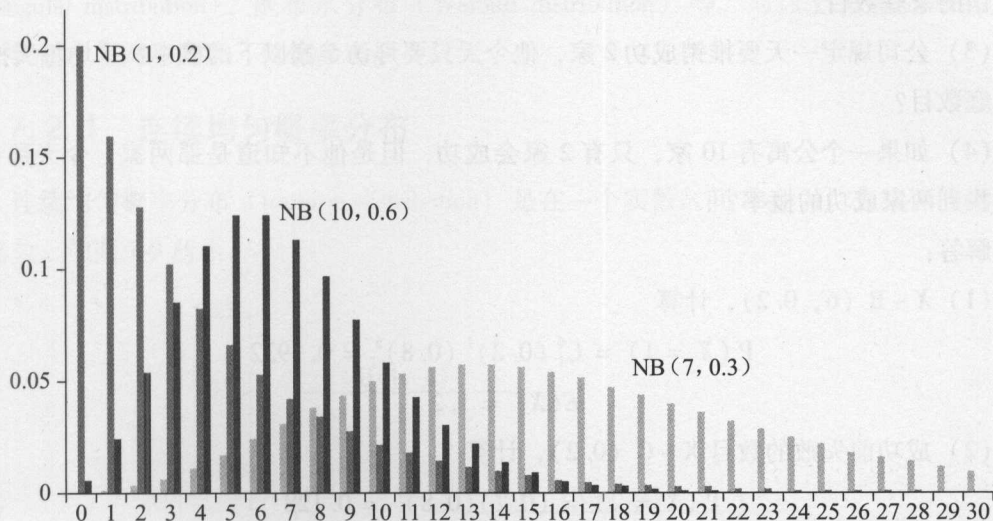


图 7-8 负二项分布

(4) 均值

$$E(X) = \frac{kq}{p}$$

(5) 方差

$$V(X) = \frac{kq}{p^2}$$

(6) 偏度系数

$$S(X) = \frac{1+q}{\sqrt{kq}} > 0 \text{ (右偏型)}$$

(7) 峰度系数

$$K(X) = 3 + \frac{p^2 + 6q}{kq}$$

(8) 若 $k=1$, 则负二项分布为几何分布。

(9) 独立的几何分布相加以后是负二项分布: 若 $X_i \sim G(p), i = 1, 2, \dots, n$, 且 X_i 为独立的, 则

$$Y = X_1 + X_2 + \dots + X_n \sim NB(n, p)$$

例题 7.2 推销员问题

推销员在 1 个家庭推销吸尘器成功的概率为 0.2, 请问:

(1) 他每天拜访 6 家, 只有 1 家成功的概率? 平均每天推销成功的数目?

(2) 公司规定一天要推销成功 1 家, 他今天刚好拜访 3 家就完成任务的概率? 平均每天推销的家庭数目?

(3) 公司规定一天要推销成功 2 家, 他今天只要拜访 5 家以下的概率? 平均每天推销的家庭数目?

(4) 如果一个公寓有 10 家, 只有 2 家会成功, 但是他不知道是那两家, 今天拜访 5 家, 找到两家成功的概率?

解答:

(1) $X \sim B(6, 0.2)$, 计算

$$P(X = 1) = C_1^6 (0.2)^1 (0.8)^5 = 0.3932$$

$$E(X) = 1.2$$

(2) 成功前失败的数目 $X \sim G(0.2)$, 计算

$$P(X = 2) = (0.2)(0.8)^2 = 0.128$$

$$E(X) = 0.8/0.2 = 4$$

要再加上最后成功的一家, 平均每天推销 5 家。

(3) 完成前失败的数目 $X \sim NB(2, 0.2)$, 计算

$$P(X \leq 3) = P(0) + P(1) + P(2) + P(3)$$

$$= (0.2)^2 + \binom{2}{1}(0.2)^2(0.8) + \binom{3}{2}(0.2)^2(0.8)^2$$

$$+ \binom{4}{3}(0.2)^2(0.8)^3$$

$$= 0.2627$$

$$E(X) = (2)0.8/0.2 = 8$$

再加上成功的 2 家, 平均每天推销 10 家。

(4) $X \sim HG(10, 2, 5)$, 计算

$$P(X = 2) = \frac{C_2^2 C_3^8}{C_5^{10}} = \frac{56}{252}$$

7.2 连续型随机变量的概率分布

以下我们要介绍的连续型随机变量的概率密度函数有: 连续均匀概率分布、正态分

布、指数分布、t分布、卡方分布、F分布。至于对数正态分布 (logarithmic normal distribution)、贝他分布 (Beta distribution)、伽玛分布 (Gamma distribution)、三角形分布 (Trangular distribution)、威布尔分布 (Weibuli distribution) 等, 与以上概率分布的关联性, 我们放在网络资源, 仅供参考。

7.2.1 连续均匀概率分布

连续均匀概率分布 (uniform distribution) 是在一个实数区间 $[a, b]$ 有相等的概率密度函数, 如图 7-9 所示。

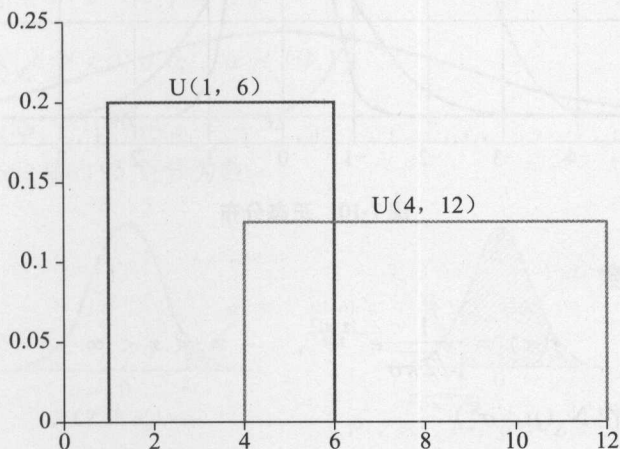


图 7-9 连续均匀概率分布

(1) 概率密度函数

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

(2) 连续均匀概率分布记作 $U(a, b)$

(3) 参数 $= a, b$

(4) 均值: $E(X) = (a+b)/2$

(5) 方差: $V(X) = (b-a)^2/12$

(6) 偏度系数: $S(X) = 0$ (对称型)。

(7) 峰度系数: $K(X) = 1.8$ (平峰型, 常数和 a, b 无关)。

(8) 利用连续均匀概率分布 $U(0, 1)$, 可以模拟出其他各种概率分布的随机变量值。

7.2.2 正态分布

正态分布 (normal distribution) 是统计学最常用的概率分布, 也是推论统计的主角。

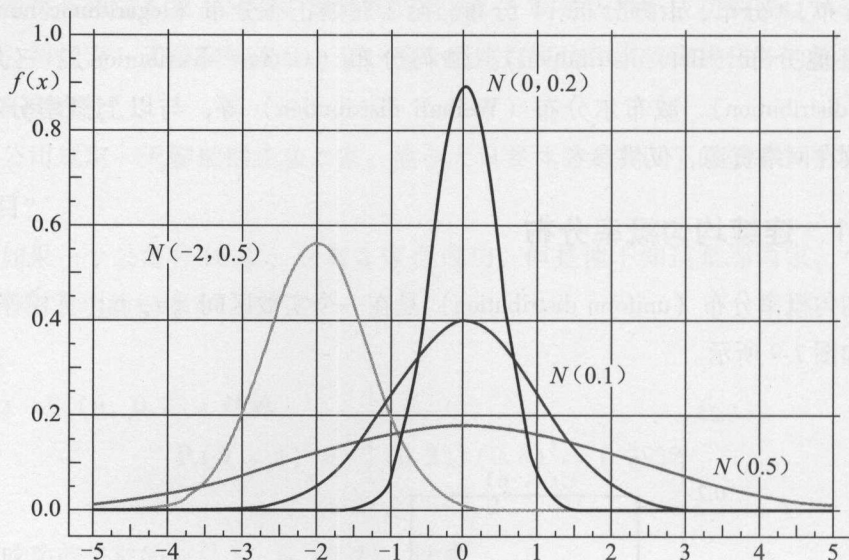


图 7-10 正态分布

1) 概率密度函数

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

2) 正态分布记作 $N(\mu, \sigma^2)$ 3) 参数 $= \mu, \mu \in \mathbb{R}, \sigma > 0$ 4) 均值: $E(X) = \mu$ 5) 方差: $V(X) = \sigma^2$ 6) 众数: $M(X) = \mu$ 7) 偏度系数: $S(X) = 0$ 8) 峰度系数: $K(X) = 3$

(正态峰, 常数和 σ^2 无关)。

9) 独立的正态分布是线性可加性, 即

若 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, 且 X 和 Y 为独立, 则 $aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$

10) 以下是正态分布在几个重要区间的概率:

(1) 正态分布在 $\mu - 0.6745\sigma, \mu + 0.6745\sigma$ 的概率为 0.5。

(2) 正态分布在 $\mu - 1.000\sigma, \mu + 1.000\sigma$ 的概率为 0.6827。

(3) 正态分布在 $\mu - 1.6450\sigma, \mu + 1.6450\sigma$ 的概率为 0.9。

(4) 正态分布在 $\mu - 1.9600\sigma, \mu + 1.9600\sigma$ 的概率为 0.95。

- (5) 正态分布在 $\mu - 2.0000\sigma, \mu + 2.0000\sigma$ 的概率为 0.9544。
 (6) 正态分布在 $\mu - 2.3260\sigma, \mu + 2.3260\sigma$ 的概率为 0.98。
 (7) 正态分布在 $\mu - 2.5760\sigma, \mu + 2.5760\sigma$ 的概率为 0.99。
 (8) 正态分布在 $\mu - 3.0000\sigma, \mu + 3.0000\sigma$ 的概率为 0.9973。
 11) 正态分布曲线的反曲点在 $\mu \pm \sigma$ 。
 12) 正态分布 $X \sim N(\mu, \sigma^2)$, 则 $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$, $N(0, 1)$ 是标准正态分布。

z_α 是 $N(0, 1)$ 的 $(1 - \alpha) \times 100$ 百分位数, 是统计估计与检验的一个重要符号。

有的统计学书定义 $P(Z < z_\alpha) = \alpha$ 。

但是本书定义 $P(Z > z_\alpha) = \alpha, \alpha \in (0, 1)$

$$z_{0.05} = 1.645, z_{0.025} = 1.96, z_{0.01} = 2.33$$

$z_{0.05}$ 是标准正态分布的 95 百分位数。

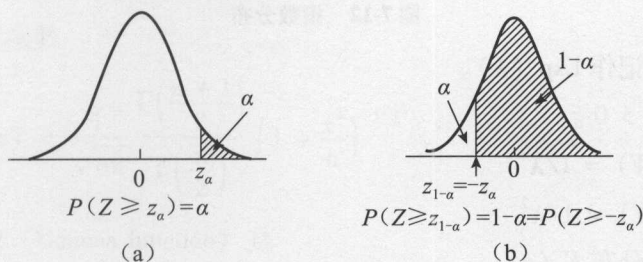


图 7-11 z_α 的定义

(a) $Z_\alpha > 0$; (b) $Z_{1-\alpha} < 0$

13) 二项分布与泊松分布都可以利用正态分布, 计算其近似的概率。

14) 一般正态分布 $N(\mu, \sigma^2)$ 的 k 百分位数 P_k , 计算

$$P_k = \mu + z_{1-k/100} \times \sigma$$

7.2.3 指数分布

指数分布 (exponential distribution) 通常适用在相邻两个发生事件的时间, 例如:

- (1) 某超市顾客到达的间隔时间。
- (2) 工厂机器故障的间隔时间。
- (3) 打进电话的间隔时间。

指数分布如图 7-12 所示。

(1) 概率密度函数

$$f(x) = \lambda e^{-\lambda x}, x > 0$$

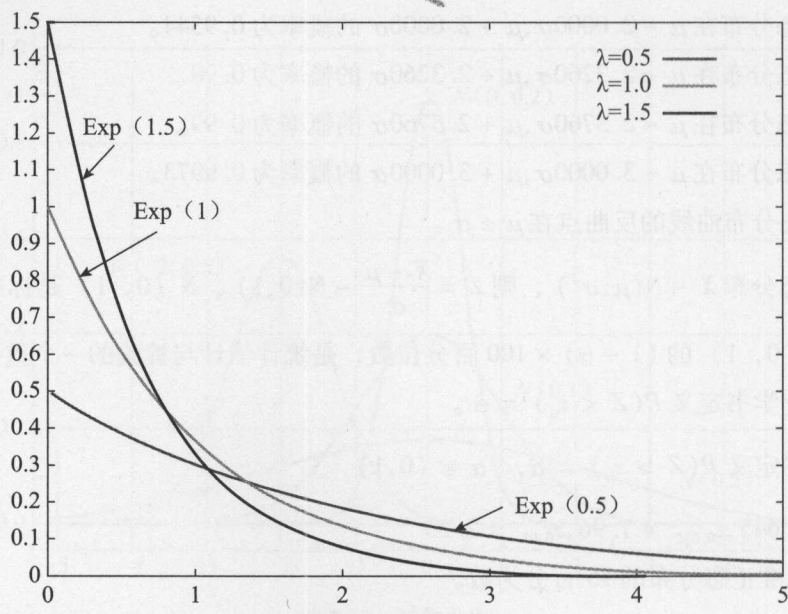


图 7-12 指数分布

- (2) 指数分布记作 $\text{Exp}(\lambda)$ 。
- (3) 参数 $= \lambda > 0$ 。
- (4) 均值: $E(V) = 1/\lambda$
- (5) 方差: $V(V) = 1/\lambda^2$
- (6) 累积概率分布 $F(x)$

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}$$

$$P(X \geq x) = e^{-\lambda x}$$

- (7) 偏度系数: $S(X) = 2$ (非常右偏型, 常数与概率分布的参数 λ 无关)。
- (8) 峰度系数: $K(X) = 9$ (非常尖峰型, 常数与概率分布的参数 λ 无关)。
- (9) 指数分布在连续型概率分布中唯一具有“无记忆性质” (memoryless property)。

以下是连续型无记忆性的数学定义:

$$P(X > s + t | X > t) = P(X > s), \text{ 对所有的 } s \geq 0, t \geq 0 \text{ 均成立。}$$

- (10) 若 $X \sim \text{Exp}(\lambda)$, 则 cX 也是指数分布, 参数 λ/c , $cX \sim \text{Exp}(\lambda/c)$ 。

(11) 若单位时间 (例如一小时) 内到达的人数是泊松分布 $\text{Pois}(\lambda)$, 则相邻到达者的间隔时间 (前后两个人到达的距离时间) 是指数分布 $\text{Exp}(\lambda)$ 。

7.2.4 t 分布

t 分布 (t distribution) 运用在: 当总体方差 σ^2 未知时, 总体均值的区间估计和检验。t 分布如图 7-13 所示。

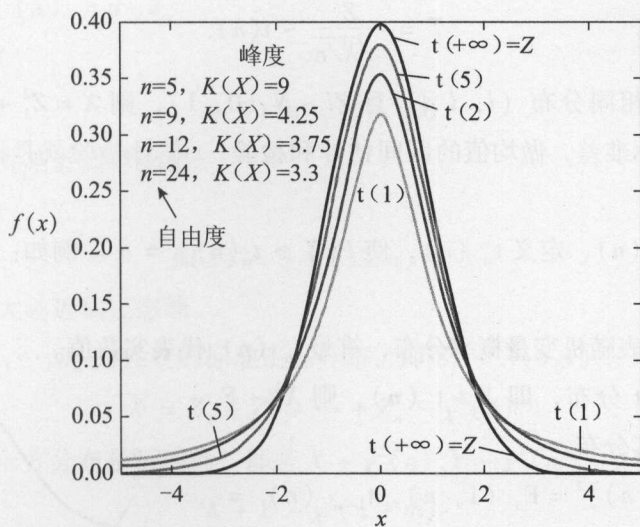


图 7-13 t 分布

(1) 概率密度函数

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}, \quad -\infty < x < +\infty$$

(2) 伽玛函数 (Gamma function) 是

$$\Gamma(\alpha) = \int_0^{\infty} e^{-y} y^{\alpha-1} dy$$

(3) 参数 = n (正整数), 称为 t 分布的自由度, 记作 $t(n)$ 。

(4) 均值: $E(X) = 0$

(5) 方差

$$V(X) = \frac{n}{n-2}, n > 2$$

(6) 众数: $M(X) = 0$

(7) 偏度系数: $S(X) = 0$

(8) 峰度系数

$$K(X) = \frac{3(n-2)}{n-4} > 3 \quad (n > 4, \text{尖峰型})$$

虽然 $t(1), \dots, t(4)$, 没有峰度系数, 但 $t(1)$ 分布中间尖窄, 两边尾巴厚长, 尖峰型。

(9) 若 $Z \sim N(0,1)$, $X \sim \chi^2(n)$, 则

$$T = \frac{Z}{\sqrt{X/n}} \sim t(n)$$

(若 Z_i 是独立相同分布 ($i, i. d$) 且 $Z_i \sim N(0, 1)$), 则 $X = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi^2(n)$, 所以用抽样标准差, 做均值的区间估计和检验, t 估计检验就是根据这个性质, 请见 9.5 节)

(10) 若 $X \sim t(n)$, 定义 $t_\alpha(n)$, 使 $P[X > t_\alpha(n)] = \alpha$ 。例如: $\alpha = 0.01, n = 20$, $t_\alpha(n) = 1.325$ 。

符号 $t(n)$ 代表随机变量概率分布, 符号 $t_\alpha(n)$ 代表实数值。

(11) 若 X 为 t 分布, 即 $X \sim t(n)$, 则 $X^2 \sim F(1, n)$, 即 X^2 是 F 分布。

(12) $[t_{\alpha/2}(n)]^2 = F_\alpha(1, n)$, $t_{1-\alpha}(n) = -t_\alpha(n)$, 如图 7-14 所示。

(13) 若 $n \rightarrow \infty$, 则 $t(n) \rightarrow Z = N(0, 1)$ 。

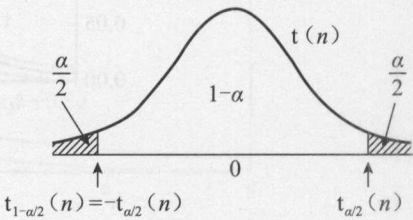


图 7-14 $t_{\alpha/2}$ 的定义

7.2.5 卡方分布

卡方分布 (chi-square distribution) 通常运用在: 总体方差 σ^2 的估计和检验、分类数据卡方检验, 以及非参数统计的 KW 检验, 卡方分布如图 7-15 所示。

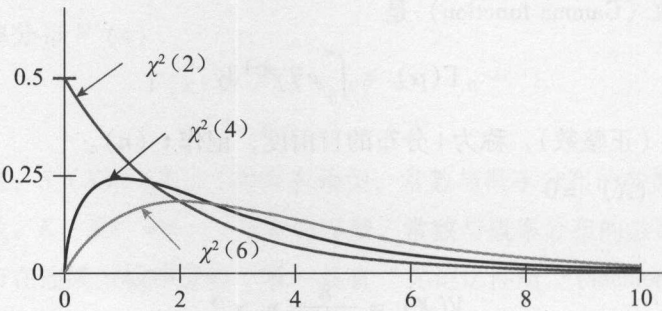


图 7-15 卡方分布

(1) 概率密度函数

$$f(x) = \frac{x^{(n-2)/2} \exp(-x/2)}{2^{n/2} \Gamma(n/2)}, x > 0$$

(2) 参数 $= n$, 正整数, 称为卡方分布的自由度 (degree of freedom)。

(3) 随机变数 X 为自由度 n 的卡方分布, 记作 $X \sim \chi^2(n)$ 。

(4) 均值: $E(X) = n$

(5) 方差: $V(X) = 2n$

(6) 众数: $M(X) = n - 2$

(7) 偏度系数

$$S(X) = \sqrt{8/n} > 0$$

右偏型, n 越大越近似对称型。

(8) 峰度系数

$$K(X) = 3 + (12/n) > 3$$

尖峰型, n 越大越近似正态峰。

(9) 若 Z_1, Z_2, \dots, Z_n 为独立的标准正态分布, 即 $Z_i \sim N(0, 1)$, 则

$$X = Z_1^2 + Z_2^2 + \dots + Z_n^2 \sim \chi^2(n)$$

(10) 独立的卡方分布是可加性, 即若 $X \sim \chi^2(n), Y \sim \chi^2(m)$, 且 X 和 Y 为独立, 则

$$X + Y \sim \chi^2(n + m)$$

(11) 若 $X \sim \chi^2(n)$, 定义 $\chi_\alpha^2(n)$, 使 $P[X > \chi_\alpha^2(n)] = \alpha$ 。

例如: $\alpha = 0.05, n = 10, \chi_{\alpha, n}^2 = 18.307$ 。(利用查表或中文统计)

请注意: 符号 $\chi^2(n)$ 代表随机变量概率分布, 符号 $\chi_\alpha^2(n)$ 代表实数值。

$$(12) \chi_\alpha^2(1) = (z_{\alpha/2})^2$$

如图 7-16 所示。

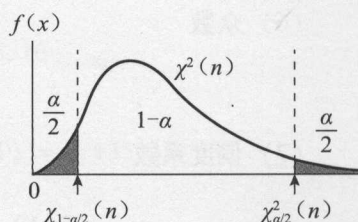


图 7-16 $\chi_{\alpha/2}^2(n)$ 的定义

7.2.6 F 分布

F 分布 (F distribution) 通常运用在: 两个总体方差比 σ_1^2/σ_2^2 的区间估计和检验、方差分析、回归分析, 以及贝努里分布总体比例值小样本之区间估计 (请见补充教材)。F 分布如图 7-17 所示。

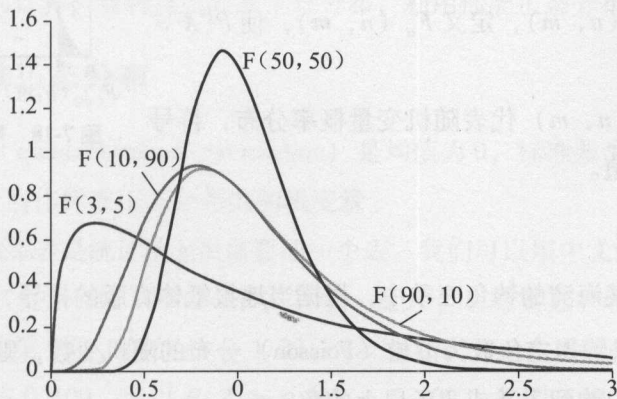


图 7-17 F 分布

(1) 概率密度函数

$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)n^{\frac{n}{2}}m^{\frac{m}{2}}x^{\frac{n-2}{2}}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)(m+nx)^{\frac{n+m}{2}}}, \quad x > 0$$

(2) F 分布记作 $F(n, m)$ 。

(3) 参数 $= n, m$, 正整数, 为 F 分布的自由度, n 是分子的自由度, m 是分母的自由度。

(4) 均值

$$E(X) = \frac{m}{m-2}, \quad m > 2$$

(5) 方差

$$V(X) = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}, m > 4$$

(6) 众数

$$M(X) = \frac{m(n-2)}{n(m+2)}$$

(7) 偏度系数

$$S(X) = \frac{(2n+m-2)\sqrt{8(m-4)}}{(m-6)\sqrt{n^2+nm-2n}}, m > 6$$

(8) 峰度

$$K(X) = 3 + \frac{12[(m-2)^2(m-4) + n(n+m-2)(5m-22)]}{n(m-6)(m-8)(n+m-2)}, m > 8 \text{ (尖峰型)}$$

(9) 若 X, Y 分别为自由度 n, m 的卡方分布, $X \sim \chi_n^2, Y \sim \chi_m^2$, 则 $F = \frac{X/n}{Y/m} \sim F(n, m)$ 。

(10) 若 $X \sim F(n, m)$, 定义 $F_\alpha(n, m)$, 使 $P[X > F_\alpha(n, m)] = \alpha$ 。

注意: 符号 $F(n, m)$ 代表随机变量概率分布, 符号 $F_\alpha(n, m)$ 代表实数值。

如图 7-18 所示。

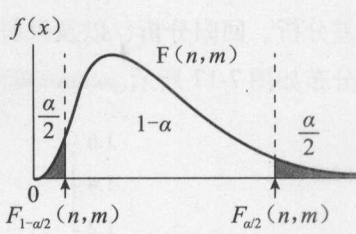


图 7-18 $F_{\alpha/2}(n, m)$ 的定义

例题 7.3 在某海湾的钓鱼旺季中, 根据当地报纸体育版的报导, 钓鱼者平均每小时可钓到两条鱼。如果钓得之鱼数为泊松 (Poisson) 分布的随机变数, 则下列概率为何:

(1) 一小时之内钓到 3 条或 3 条以上的鱼?

- (2) 半小时之内至少钓到 3 条鱼?
 (3) 4 小时之内大于 10 条鱼?
 (4) 2 小时以上没钓到鱼之概率?
 (5) 已经有 1 小时没钓到鱼, 则接下来一小时仍没钓到鱼之概率?
 (6) 已经有 5 小时没钓到鱼, 则接下来一小时仍没钓到鱼之概率?

解答:

- (1) 一小时内钓到鱼的数目 $X \sim \text{Pois}(2)$, 计算

$$P(X \geq 3) = 1 - 0.6767$$

- (2) 半小时内钓到鱼的数目 $X \sim \text{Pois}(1)$, 计算

$$P(X \geq 3) = 1 - 0.9197$$

- (3) 4 小时内钓到鱼的数目 $X \sim \text{Pois}(8)$, 计算

$$P(X > 10) = 1 - 0.8159$$

- (4) 钓到鱼的间隔时间 $T \sim \text{Exp}(2)$, 计算

$$P(T > 2) = e^{(-2) \times 2} = e^{(-4)} = 0.018316$$

- (5) $T \sim \text{Exp}(2)$, 计算

$$P(T > 2 | T > 1) = e^{-4}/e^{-2} = e^{-2} = P(T > 1) = 0.135335$$

- (6) $T \sim \text{Exp}(2)$, 计算

$$P(T > 6 | T > 5) = e^{-12}/e^{-10} = e^{-2} = P(T > 1) = 0.135335$$

指数分布满足“无记忆性质”。

7.3 正态分布概率的计算

正态分布概率的计算, 要转换成标准正态分布, 利用标准正态分布表, 查出其概率。

7.3.1 标准正态分布

标准正态分布 (standard normal distribution) 是均值为 0, 标准差为 1 的正态分布, 记作 $N(0, 1)$, 以 Z 当作标准正态分布的随机变量。

标准正态分布概率表是统计学上最重要的一个表。我们可以用中文统计, 或在附表 A-1 的标准正态分布概率表中利用 $P(Z \geq z)$, 其中 z 是大于 0 的实数。表的左边到小数第 1 位, 表的上方是小数第 2 位, 例如: $z = 1.23$, $P(Z \geq 1.23)$, 在表的左边查 1.2, 在上方查 0.03, 两者交会点 0.109, 所以 $P(Z \geq 1.23) = 0.109$, 如图 7-19 (a) 所示。

另外，我们要知道 z 是多少，使 $P(Z \geq z) = a$ ，其中 a 已知。例如： $a = 0.025$ ，在表 A1 中，找到 0.025，往外推出得到 $z = 1.96$ ，如图 7-19 (b) 所示。

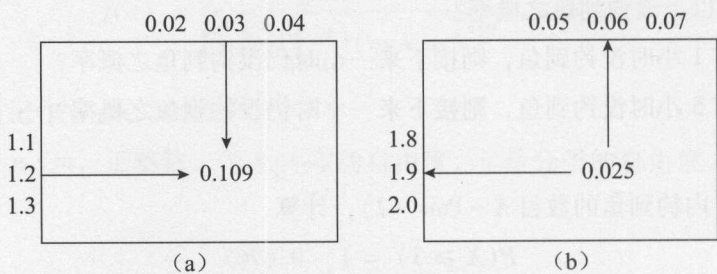


图 7-19 标准正态分布查表法

(a) 方法 1; (b) 方法 2

7.3.2 一般正态分布

一般正态分布的概率计算，要经过标准化转换，变成标准正态分布。 X 是正态分布，其均值 μ ，标准差 σ ， $X \sim N(\mu, \sigma^2)$ 。求 $P(X \geq x)$ ，标准化转换如下：

标准正态分布 $Z = \frac{X - \mu}{\sigma}$ ；标准分数 $z = \frac{x - \mu}{\sigma}$ ，如图 7-20 所示。

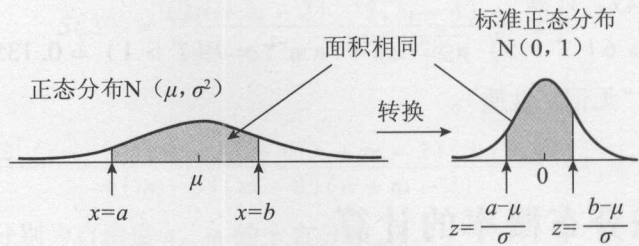


图 7-20 转换前后，面积相同

7.3.3 利用正态分布求二项分布的概率

利用正态分布求二项分布的概率，假设条件与步骤：

- 1) 若二项分布 $B(n, p)$ ， $np > 5$ ， $n(1 - p) > 5$ ，则可利用正态分布求其近似值。
- 2) 利用正态分布 $N(\mu, \sigma^2)$ ， $\mu = np$ ， $\sigma = \sqrt{np(1 - p)}$ 。
- 3) 因为二项分布是离散型，正态分布是连续型，所以计算区间概率，要修正其上下限。下限减 0.5，上限加 0.5。这个步骤称为连续性修正 (continuity correction)。

例如： $X \sim B(n, p)$ 是二项分布， $Y \sim N[np, np(1 - p)]$ ，求

- (1) $P(3 \leq X \leq 8) \cong P(2.5 \leq Y \leq 8.5)$
 (2) $P(0 < X < 6) = P(1 \leq X \leq 5) \cong P(0.5 \leq Y \leq 5.5)$
 (3) $P(X = 5) \cong P(4.5 \leq Y \leq 5.5)$

如果 n 很大, 而使 $\sqrt{np(1-p)}$ 也很大, 则可以不必要做连续性修正。

例题 7.4—7.12 (见网络资源)

7.4 中文统计应用

7.4.1 离散型概率分布——二项分布 (例题 7.2)

执行“中文统计”→“概率分布”→“离散型概率分布”→“二项分布”命令。在浅绿色的单元格输入数据, 离开单元格, 单击“重新计算”。操作示意图如图 7-21 所示。

6. 概率分布	离散型概率分布
7. 抽样理论	连续型概率分布
8. 区间估计(一个总体)	正态分布
9. 假设检验(一个总体)	概率分布查表

1. 二项分布 概率值					2. 二项分布概率值 -- 利用正态分布近似				
输入数据					输入数据				
试验的次数(n)	n =	20.00			试验的次数	n =	20		
试验成功的概率(p)	p =	0.80			试验成功的概率	p =	0.8		
试验成功的次数(i)	i =	17.00			试验成功的次数	i =	17		
					求区间 (i1	14.00	18.00	=i2) 之概率	
输出数据					输出数据				
二项分布概率值 $P(X = i) = 0.2054$					无连续性修正 连续性修正				
二项分布累积概率值 $P(X \leq i) = 0.7939$					概率值 $P(X = i) = 0 0.1891$				
概率值 $P(X \geq i) = 0.4114$					累积概率值 $P(X \leq i) = 0.7119 0.7991$				
					概率值 $P(X \geq i) = 0.2881 0.3899$				
					概率值 $P(i1 \leq X \leq i2) = 0.7364 0.8377$				
均值	方差	标准差	偏态	峰态					
16.0000	3.2000	1.7889	(0.3354)	3.0125					
二项分布					泊松分布 超几何分布 负二项分布 几何分布				

图 7-21 操作示意图

7.4.2 连续型概率分布——正态分布 (例题 7.4)

执行“正态分布”的操作示意图如图 7-22 所示。

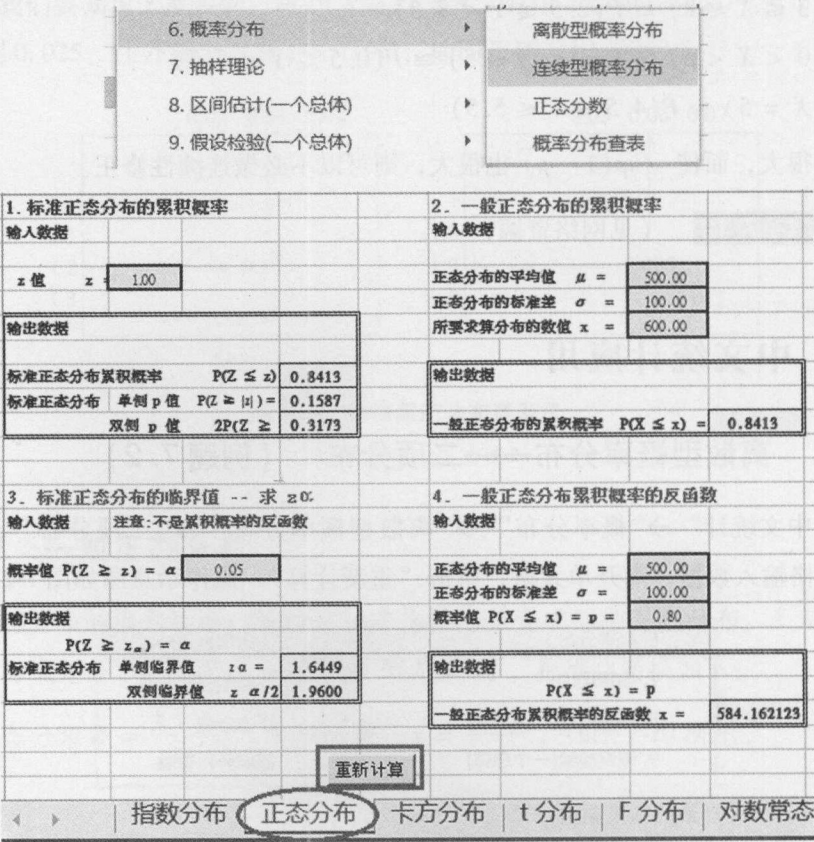


图 7-22 执行“正态分布概率”的操作示意图

7.4.3 连续型概率分布——卡方分布（例题 7.5）

执行“卡方分布”的操作示意图如图 7-23 所示。

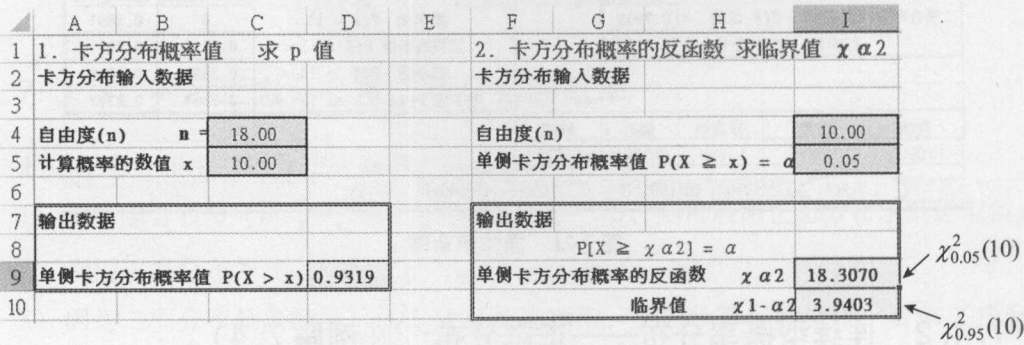


图 7-23 执行“卡方分布概率”的操作示意图

7.4.4 连续型概率分布——t 分布（例题 7.6）

执行“t 分布”的操作示意图如图 7-24 所示。

	A	B	C	D	E	F	G	H	I
1	1. t 分布概率值 求 p 值					2. t 分布概率的反函数 求 $t_{\alpha/2}$, t_{α} 临界值			
2	输入数据					输入数据			
3									
4	自由度(n)	n =	10.00			自由度(n)	n =	10.00	
5	计算 t 分布概率的数值	x =	2.00			显著水平		0.05	
6									
7	输出数据					输出数据			
8						$P[X > t_{\alpha}] = \alpha$			
9	t 分布单侧概率值	$P(X > x) =$	0.0367			双侧 t 分布概率的反函数	$t_{\alpha/2}$	2.2281	$t_{0.025}(10)$
10	t 分布双侧概率值	$P(X > x \text{ OR } X < -x) =$	0.0734			单侧 t 分布概率的反函数	t_{α}	1.8125	$t_{0.05}(10)$

图 7-24 执行“t 分布概率”的操作示意图

7.5 本章思维导图

本章思维导图如图 7-25 所示。

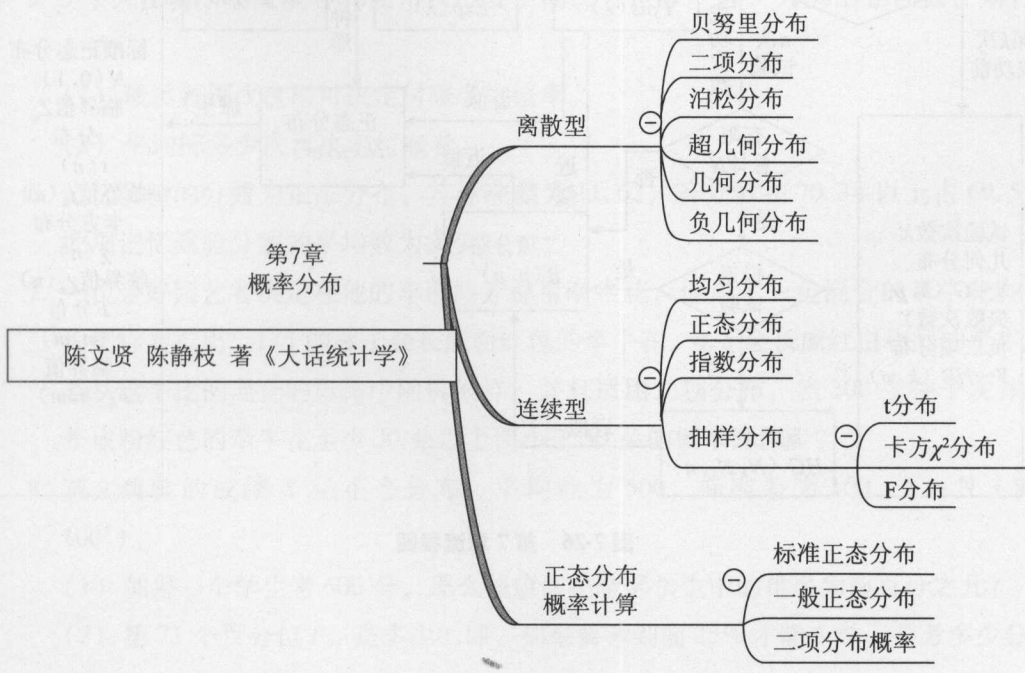


图 7-25 第 7 章思维导图

习 题

- 公交车到达时间为 10:00a. m. 到 10:30a. m. 的连续均匀分配。
 - 公交车在 10:10a. m. 以后到达之概率?
 - 公交车在 10:20a. m. 以前到达之概率?
 - 等到 10:15a. m. 还没有公交车, 则至少再等 10min 以上之概率?
- 是非题以投掷铜板方式来作答, 若正面则答“是”, 若反面则答“否”。是非题共有 100 题, 每题 1 分, 请问至少有 60 分的概率为多少?
- 一个骰子投掷 180 次, 记录出现的点数, 则下列概率为何?
 - 6 点出现 20 次的概率?
 - 奇数点出现 40 次以下的概率?
 - 3 的倍数点出现至少 50 次的概率?
- 可乐公司的每 6 罐可乐中, 有一个免费再来一罐的拉环, 如果要有 95% 的概率, 得到免费一罐的拉环, 要买多少罐可乐?
- 3 个人在餐厅吃完饭各掷硬币, 出现不同面者付账。若 3 人掷出相同面, 则再掷一次。
 - 最多各掷 3 次即可决定付账者之概率。
 - 平均掷多少次可决定付账者。
- 记忆测验的分数为正态分布, 其标准差为 11.62, 若分数在 70.34 以上占 69.5%, 此项记忆测验分数的平均数为多少?
- 一位爱好园艺者决定在他的车道两旁种植牵牛花, 他买了一包混合的牵牛花种子, 根据标签指出, $1/5$ 的种子会长成粉红色的牵牛花, $4/5$ 会长成红白相间的牵牛花。若从这个比例混合的母体中随机抽样, 并且适用二项分布, 当 200 个种子发芽时, 长成粉红色的牵牛花至少 30 朵以上而小于 50 朵的概率为多少?
- 英文测验的成绩 X 是正态分布, 平均数为 500, 标准差为 100, $X \sim N(500, 100^2)$ 。
 - 如果一个学生考 600 分, 那么他应该在全部学生中的排名为前百分之几?
 - 第 75 个百分位 P_{75} 是多少? 即, 如果要考到前 25% 才能入学, 要考多少分?

其他习题请下载。



第 8 章

抽 样 理 论

任凭弱水三千，我只取一瓢饮。

——曹雪芹《红楼梦》

“数大”就（便）是美。

——徐志摩《西湖记》

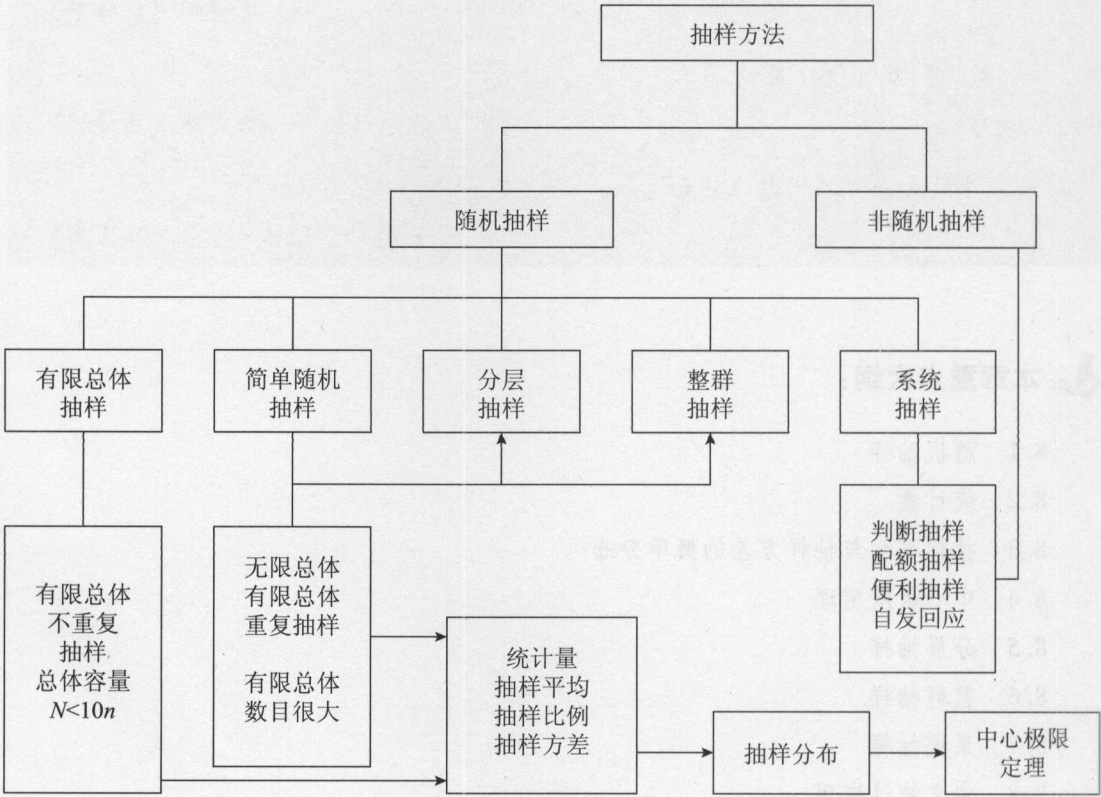
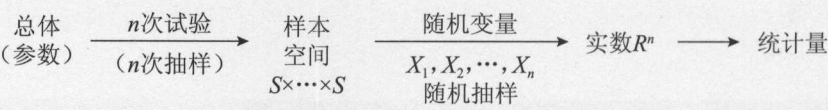
以小致大，谓之“抛砖引玉”。

——《幼学琼林·卷三·珍宝类》



本章重点大纲：

- 8.1 随机抽样
- 8.2 统计量
- 8.3 抽样平均与抽样方差的概率分布
- 8.4 中心极限定理
- 8.5 分层抽样
- 8.6 整群抽样
- 8.7 系统抽样
- 8.8 中文统计应用
- 8.9 本章流程图
- 8.10 本章思维导图



本章概念图

8.1 随机抽样

抽样是统计学的一个重要步骤,为了收集数据,除了现成的二手数据,不管是调查数据(问卷调查、电话访问、选举的出口民调、产品抽验、路边访谈等),或者是实验数据、观察数据的收集,都需要抽样。

本章除了介绍抽样方法,也说明抽样统计量的概率分布、期望值、方差,可以对照第2章叙述统计的一些量数,这也是以后各章推论统计的基础。

定义 有限总体 (finite population) 是指总体的样本点的数目是有限个。

定义 无限总体 (infinite population) 是指总体的样本点的数目是无限多。

若总体的随机变量为 X , 每次抽样是一个随机变量 X_i , 其结果是实数 x_i 。以下我们用 X_i 代表抽样的随机变量, x_i 代表抽样的实际值 (抽样值)。

定义 总体是有限总体, 如果每次抽样每个样本点的概率均等, 而且每次抽样后, 样本又放回总体, 称为重复抽样 (sampling with replacement); 如果每次抽样每个样本点的概率均等, 但是每次抽样后, 样本不放回总体, 称为不重复抽样 (sampling without replacement)。

定义 若 n 次抽样 X_1, X_2, \dots, X_n , 满足下列条件, 则称为非常简单随机抽样 (very simple random sample, 简写 VSRS), 简称随机抽样 (random sample)。

- (1) X_1, X_2, \dots, X_n 为独立。
- (2) X_1, X_2, \dots, X_n 的概率分布, 与总体随机变量 X 的概率分布相同。

定义 如果总体是有限总体, 每次抽样为“不重复式抽样”, 则 n 次抽样 X_1, X_2, \dots, X_n 不是独立的; X_1, X_2, \dots, X_n 的概率分布不相同。这种抽样称为有限总体随机抽样 (finite population random sample)。

以下3种情况是(非常简单)随机抽样。

- (1) 如果总体是无限总体, 每次抽样每个样本点的概率均等, 则为随机抽样。
- (2) 如果总体是有限总体, 每次抽样为“重复抽样”, 则为随机抽样。
- (3) 如果总体数目 N 大于样本量目很多 ($N > 10n$), 有限总体随机抽样近似随机抽样。

我们进行统计推论, 都是以“随机抽样”为主。

统计推论之主要目的，是要了解总体随机变量的概率分布。因为总体随机变量的概率分布是尚未完全定义的，或者说总体随机变量并不清楚其分布情形。有可能只知道其概率分布（例如：只知道是正态分布或泊松分布），但是参数未知；也有可能连其概率分布都不知道，只晓得是连续型或间断型。

我们假设总体是正态分布或近似正态分布（利用中心极限定理）去做统计推论，称作“有参数推论”（parametric inference）。如果没有假设总体随机变量的概率分布，所做统计推论，称作“非参数推论”（nonparametric inference）。

本章是有参数统计的前身，就是先假设总体分布（例如正态分布），然后推导出“抽样统计量”的概率分布。第9章开始，就根据这些抽样统计量的概率分布，定出估计与检验的推论法则。所以本章是总体特性（分布）已知、演绎抽样统计量的分布。

例题 8.1 随机抽样。（解答见网络资源）

8.2 统计量

定义 X_1, X_2, \dots, X_n 为一个随机抽样，若 $f(X_1, X_2, \dots, X_n)$ 为一个 X_1, X_2, \dots, X_n 的函数，而不包括总体分布的未知参数，则 $f(X_1, X_2, \dots, X_n)$ 称作一个统计量（statistic）。

例如： $f(X_1, X_2, \dots, X_n) = \frac{\text{Min}X_i + \text{Max}X_i}{2}$ ，即 X_i 的最小值与最大值的平均是一个统计量。例如： $f(X_1, X_2, \dots, X_n) = \sum_{X_i < 0} \text{rank}|X_i|$ ， $\text{rank}|X_i|$ 是 X_i 的绝对值的排序，取其原来是负的 X_i 的排序相加，是一个统计量，用在非参数统计。

定义 统计量 $\bar{X} : \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$ ，称作抽样平均（sample mean）。

统计量 $(w_1X_1 + w_2X_2 + \dots + w_nX_n) / \sum w_i$ ，称作加权抽样平均（sample mean）。

定义 统计量 $S^2 : S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ 称作抽样方差（sample variance）。

请将 S^2 当作一个符号，随机变量的符号，而不要当作 S 的平方。

定义 统计量 $S : S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$ ，称作抽样标准差（sample standard deviation）。

统计量是一个随机变量，如果代入抽样值，就成为统计值，是一个实数。

“统计量”与“统计值”，针对某个总体参数，就是第9章的“估计量”与“估计值”。

在本章主要讨论的是“抽样平均”与“抽样方差”，这两个统计量的概率分布。

8.3 抽样平均与抽样方差的概率分布

8.3.1 抽样平均的期望值与方差

定理 若总体随机变量 X 的概率分布为任何概率分布， X_i 是抽样的随机变量，期望值 $E(X_i) = \mu$ ，方差 $V(X_i) = \sigma^2$ ，样本量为 n ，则

$$E(\bar{X}) = \frac{\sum_{i=1}^n E(X_i)}{n} = \frac{n\mu}{n} = \mu$$

$$V(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n V(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

随机抽样的抽样平均 \bar{X} 的概率分布的期望值为 μ ，方差为 σ^2/n 。

每个随机抽样和总体有相同的分布。因为平均的结果，当样本量越大，“抽样平均”的数据会越“集中”在平均数，方差是 σ^2/n ，这就是大数法则 (law of large number)：当抽样数目越多，抽样平均会越集中在总体平均数，第 5.4.2 节的相对次数是相同道理。当移动平均期数越大，时间数列的“移动平均”曲线会较原始数列曲线更“平滑”（消除激烈的变动，见第 3 章）。

定理 若总体为有限，总体的数目是 N ，概率分布的期望值为 μ ，方差为 σ^2 ，样本量为 n ，则：有限总体随机抽样（不重复抽样）的平均 \bar{X} 的概率分布的期望值为 μ ，方差为

$$\frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right)$$

如果 $n = N$ ，即抽样数目等于总体数目，因为不重复抽样，所以样本集合为总体，抽样平均的方差为 0，即抽样平均值等于总体平均。

如果 N 比 n 大很多，例 $N > 100n$ ，则 $\frac{N-n}{N-1} \approx 1.00$ ，所以有限总体不重复抽样等于随机抽样。

8.3.2 抽样方差的期望值

定理 若抽样方差 $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$, 则 $E(S^2) = \sigma^2$ 。

$$\text{证明: } (n-1)S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + \sum_{i=1}^n \bar{X}^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2$$

$$E[(n-1)S^2] = nE[X_i^2] - nE[\bar{X}^2]$$

$$E[X_i^2] = \mu^2 + \sigma^2, \quad E[\bar{X}^2] = \mu^2 + \sigma^2/n$$

$$E[(n-1)S^2] = (n-1)\sigma^2$$

$$E(S^2) = \sigma^2$$

所以, 随机抽样的抽样方差的概率分布的期望值为 σ^2 。

8.3.3 抽样平均与抽样方差的概率分布

定理 X_1, X_2, \dots, X_n 为一个随机抽样, 总体是正态分布, 平均数为 μ , 方差为 σ^2 。

即 X_i 是独立的, 且 $X_i \sim N(\mu, \sigma^2)$ 。则:

(1) \bar{X} 与 S^2 是独立的。

(2) \bar{X} 是正态分布, 平均数为 μ , 方差 σ^2/n : $\bar{X} \sim N(\mu, \sigma^2/n)$ 。

(3) $\frac{(n-1)S^2}{\sigma^2}$ 是卡方分布, 自由度 $n-1$: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ 。

(4) $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ 是 t 分布, 自由度 $n-1$: $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$ 。

所以: $E(S^2) = \sigma^2, V(S^2) = 2\sigma^4/(n-1)$

例题 8.2 假如总体有 5 个数据: 76, 78, 79, 81, 86。计算抽样平均 \bar{X} 的概率分布。

(1) 抽样数目 $n=2$, 有限总体随机抽样 (不重复抽样)。

(2) 抽样数目 $n=2$, 随机抽样 (重复抽样)。

(3) 抽样数目 $n=4$, 有限总体随机抽样 (不重复抽样)。

解答: 总体数据 {76, 78, 79, 81, 86} 的平均数 μ , 标准差 σ : $\mu = 80, \sigma = 3.406$ 。

(1) 抽样数目 $n=2$, 有限总体随机抽样 (抽样后不重复)。可能的随机抽样为 $C_2^5 = 10$ 种:

$\{76, 78, \bar{x} = 77.0\}$, $\{76, 79, \bar{x} = 77.5\}$, $\{76, 81, \bar{x} = 78.5\}$, $\{76, 86, \bar{x} = 81.0\}$,
 $\{78, 79, \bar{x} = 78.5\}$, $\{78, 81, \bar{x} = 79.5\}$, $\{78, 86, \bar{x} = 82.0\}$, $\{79, 81, \bar{x} = 80.5\}$,
 $\{79, 86, \bar{x} = 82.5\}$, $\{81, 86, \bar{x} = 83.5\}$

抽样平均 \bar{X} 的概率分布如表 8-1 所示。

表 8-1 抽样平均 \bar{X} 的概率分布 ($n=2$, 有限总体随机抽样)

抽样平均 \bar{x}	概率 $P(\bar{x})$	$\bar{x}P(\bar{x})$	$\bar{x}^2P(\bar{x})$
77.0	0.1	7.70	592.900
77.5	0.1	7.75	600.625
78.5	0.2	15.70	1232.450
79.5	0.1	7.95	632.025
80.0	0.1	8.00	640.000
81.0	0.1	8.10	656.100
82.0	0.1	8.10	672.400
82.5	0.1	8.25	680.625
83.5	0.1	8.35	697.225
总和 Σ	1.0	80.00	6404.350

抽样平均 \bar{X} 的期望值 $\mu_{\bar{X}}$, 标准差 $\sigma_{\bar{X}}$:

$$\mu_{\bar{X}} = 80$$
$$\sigma_{\bar{X}} = \sqrt{\sum \bar{x}^2 P(\bar{x}) - (\mu_{\bar{X}})^2} = \sqrt{6404.35 - 6400} = 2.086$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{3.406}{\sqrt{2}} \sqrt{\frac{5-2}{5-1}} = 2.086$$

(2) 抽样数目 $n=2$, 随机抽样 (重复抽样)。所有可能的随机抽样有 25 种:

$\{76, 76, \bar{x} = 76.0\}$, $\{76, 78, \bar{x} = 77.0\}$, $\{76, 79, \bar{x} = 77.5\}$, $\{76, 81, \bar{x} = 78.5\}$,
 $\{76, 86, \bar{x} = 81.0\}$, $\{78, 76, \bar{x} = 77.0\}$, $\{78, 78, \bar{x} = 78.0\}$, $\{78, 79, \bar{x} = 78.5\}$,
 $\{78, 81, \bar{x} = 79.5\}$, $\{78, 86, \bar{x} = 82.0\}$, $\{79, 76, \bar{x} = 77.5\}$, $\{79, 78, \bar{x} = 78.5\}$,
 $\{79, 79, \bar{x} = 79.0\}$, $\{79, 81, \bar{x} = 80.0\}$, $\{79, 86, \bar{x} = 82.5\}$, $\{81, 76, \bar{x} = 78.5\}$,
 $\{81, 78, \bar{x} = 79.5\}$, $\{81, 79, \bar{x} = 80.0\}$, $\{81, 81, \bar{x} = 81.0\}$, $\{81, 86, \bar{x} = 83.5\}$,
 $\{86, 76, \bar{x} = 81.0\}$, $\{86, 78, \bar{x} = 82.0\}$, $\{86, 79, \bar{x} = 82.5\}$, $\{86, 81, \bar{x} = 83.5\}$,
 $\{86, 86, \bar{x} = 86.0\}$

抽样平均 \bar{X} 的期望值 $\mu_{\bar{X}}$, 标准差 $\sigma_{\bar{X}}$:

$$\mu_{\bar{X}} = 80$$
$$\sigma_{\bar{X}} = \sqrt{\sum \bar{x}^2 P(\bar{x}) - (\mu_{\bar{X}})^2} = \sqrt{6405.8 - 6400} = 2.408$$
$$\sigma_{\bar{X}} = \sigma / \sqrt{n} = 3.406 / \sqrt{2} = 2.408$$

抽样平均 \bar{X} 的概率分布如表 8-2 所示。

表 8-2 \bar{X} 的概率分布 ($n=2$, 随机抽样)

抽样平均 \bar{x}	概率 $P(\bar{x})$	$\bar{x}P(\bar{x})$	$\bar{x}^2P(\bar{x})$
76.0	0.04	3.04	231.04
77.0	0.08	6.16	474.32
77.5	0.08	6.20	480.50
78.0	0.04	3.12	243.36
78.5	0.16	12.56	985.96
79.0	0.04	3.16	249.64
79.5	0.08	6.36	505.62
80.0	0.08	6.40	512.00
81.0	0.12	9.72	787.32
82.0	0.08	6.56	537.92
82.5	0.08	6.60	544.50
83.5	0.08	6.68	557.78
86.0	0.04	3.44	295.84
Σ	1.0	80.00	6405.80

(3) 抽样数目 $n=4$ ，有限总体随机抽样（不重复抽样）。可能的随机抽样为 5 种：

- {76, 78, 79, 81, $\bar{x} = 78.50$ }, {76, 78, 79, 86, $\bar{x} = 79.75$ },
- {76, 78, 81, 86, $\bar{x} = 80.25$ }, {76, 79, 81, 86, $\bar{x} = 80.50$ },
- {78, 79, 81, 86, $\bar{x} = 81.00$ }

抽样平均 \bar{X} 的概率分布如表 8-3 所示。

表 8-3 \bar{X} 的概率分布 ($n=4$, 有限总体随机抽样)

抽样平均 \bar{x}	概率 $P(\bar{x})$	$\bar{x}P(\bar{x})$	$\bar{x}^2P(\bar{x})$
78.50	0.2	15.70	1232.45
79.75	0.2	15.95	1272.01
80.25	0.2	16.05	1288.01
80.50	0.2	16.10	1296.05
81.00	0.2	16.20	1312.20
Σ	1.0	80.00	6400.725

抽样平均 \bar{X} 的期望值 $\mu_{\bar{X}}$ ，标准差 $\sigma_{\bar{X}}$ ：

$$\mu_{\bar{X}} = 80$$

$$\sigma_{\bar{X}} = \sqrt{\sum \bar{x}^2 P(\bar{x}) - (\mu_{\bar{X}})^2} = \sqrt{6400.725 - 6400} = 0.85$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{3.406}{\sqrt{2}} \sqrt{\frac{5-4}{5-1}} = 0.85$$

8.4 中心极限定理

中心极限定理 (central limit theorem): 如果总体概率分布不是正态分布, 期望值为 μ , 方差为 σ^2 。若样本量相当大, 则抽样平均 \bar{X} 的概率分布会近似正态分布, 期望值为 μ , 方差为 σ^2/n 。

定理 X_1, X_2, \dots, X_n 为一个随机抽样, 总体是任何概率分布, 平均数为 μ , 方差为 σ^2 。

当抽样数目 $n \rightarrow \infty$, 则 $\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$, 如图 8-1 和图 8-2 所示。

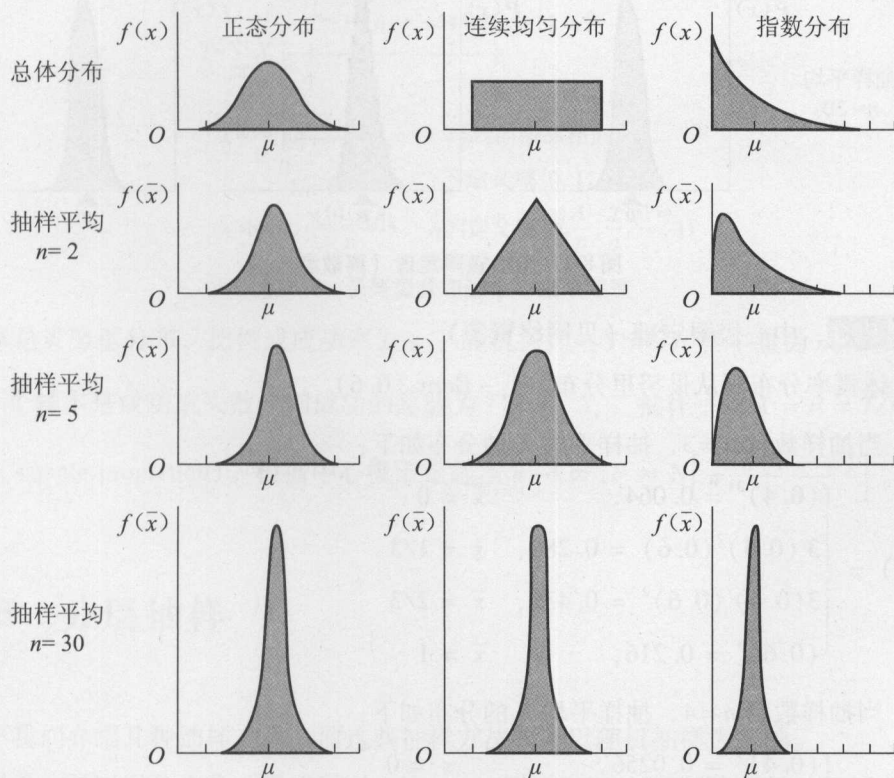


图 8-1 中心极限定理 (连续型)

依照中心极限定理, 样本量 n 相当大, 则抽样平均趋近正态分布, 但是 n 要大到什么程度, 这要看总体的概率分布。如果总体概率分布是对称的, 则 n 不必要很大, 大于 25 以上即可。如果总体概率分布是非常左偏或右偏, 则 n 要很大, 可能大于 100 以上。一般都以 $n \geq 30$ 为准则。

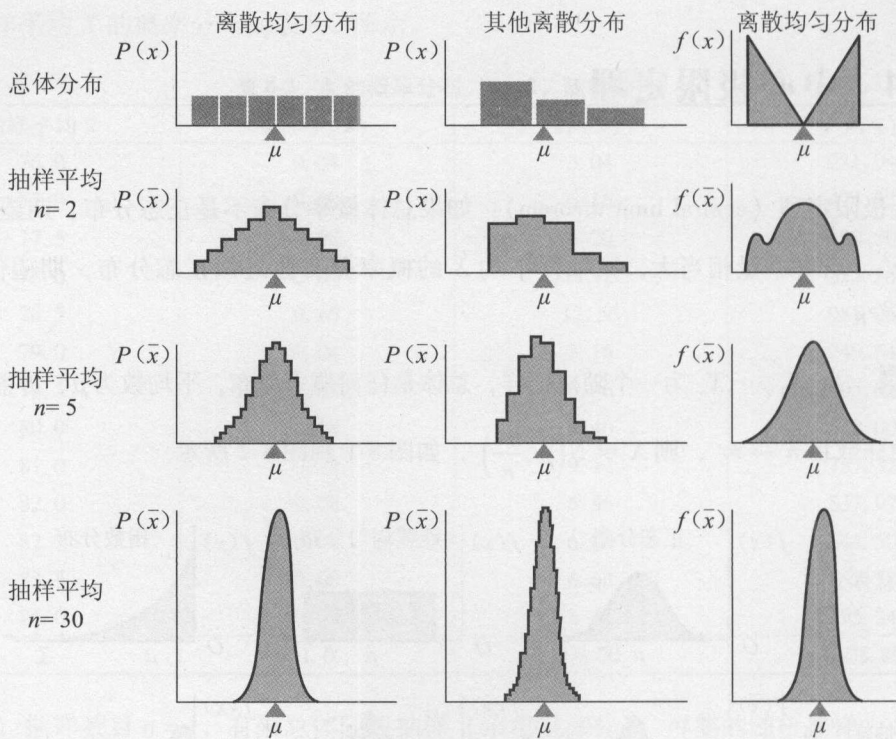


图 8-2 中心极限定理（离散型）

例题 8.3 中心极限定理（见网络资源）

若总体概率分布服从贝努里分布， $X_i \sim \text{Bern}(0.6)$

(1) 当抽样数目 $n=3$ ，抽样平均 \bar{X} 的分布如下：

$$P(\bar{x}) = \begin{cases} (0.4)^3 = 0.064, & \bar{x} = 0 \\ 3(0.4)^2(0.6) = 0.288, & \bar{x} = 1/3 \\ 3(0.4)(0.6)^2 = 0.432, & \bar{x} = 2/3 \\ (0.6)^3 = 0.216, & \bar{x} = 1 \end{cases}$$

(2) 当抽样数目 $n=4$ ，抽样平均 \bar{X} 的分布如下：

$$P(\bar{x}) = \begin{cases} (0.4)^4 = 0.0256, & \bar{x} = 0 \\ 4(0.4)^3(0.6) = 0.1536, & \bar{x} = 1/4 \\ 6(0.4)^2(0.6)^2 = 0.3456, & \bar{x} = 2/4 \\ 4(0.4)(0.6)^3 = 0.3456, & \bar{x} = 3/4 \\ (0.6)^4 = 0.1296, & \bar{x} = 1 \end{cases}$$

如果总体分布是贝努里分布 $\text{Bern}(\pi)$ ，则对任何抽样数目 n ， \bar{X} 是一个二项分布 B

(n, π) , 只是将定义范围, 从 0 到 n , 缩到 $[0, 1]$ 。所以当 n 相当大 ($n > 30$), \bar{X} 变成连续型概率分布。实际上 $n\bar{X}$ 是离散型二项分布, 所以利用正态分布来做近似, 求概率值时, 要注意连续性修正, 即计算 $P(\bar{X} \geq a)$ 改为 $P(\bar{X} \geq a - 0.5)$ 。我们知道, 如果 $np > 5$ 且 $n(1-p) > 5$, 则二项分布趋近正态分布。即

$$n\bar{X} \sim B(n, \pi) \approx N[n\pi, n\pi(1-\pi)], E(n\bar{X}) = n\pi, E(\bar{X}) = \pi$$

$$V(n\bar{X}) = n^2 V(\bar{X}) = n\pi(1-\pi), V(\bar{X}) = \frac{\pi(1-\pi)}{n}$$

所以: $\bar{X} \sim N\left[\pi, \frac{\pi(1-\pi)}{n}\right]$, 如图 8-3 所示。

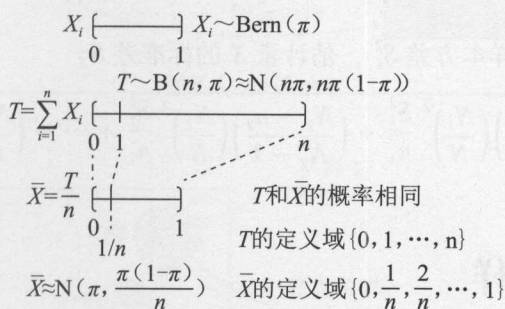


图 8-3 贝努里分布的中心极限定理

总体是贝努里分布, 比例或成功率为 π , 随机抽样 (VSRs) 样本量为 n , $X_i = 1$ 或 0, 代表第 i 个样本是成功或失败, 则成功的总数为 $T = \sum X_i$, 抽样平均 $\bar{X} = \bar{P} = T/n$ 称为样本比例 (sample proportion)。根据中心极限定理当 $n \rightarrow \infty$, $\bar{P} \approx N\left[\pi, \frac{\pi(1-\pi)}{n}\right]$ 。

8.5 分层抽样

以下我们介绍几种抽样方法, 而这些抽样方法都是以随机抽样为基础。

如果总体可以区分为几个组成层 (stratum)。例如: 以地区分层 (每个目标市场为一层, 每个选区为一层); 以职业分层 (每个职业团体为一层); 以产业分层 (每个产业为一层), 等等。

这里的“分层”是平行的层, 不是上下的阶层。

分层抽样 (stratified sampling) 又称类型抽样, 是在每一层中, 按照比例, 随机抽样找出一些个体, 组成一个样本。分层抽样的进行步骤如下。

(1) 总体中的个体依其性质、位置、种类、等级、结构等，分成 k 层。

(2) 决定抽样的个数 n 。

(3) 决定每层的抽样数目 n_i 。可利用比例方式，例如总体总数 N ，第 i 层总体数目为 N_i ，则第 i 层的抽样数目为 n_i ： $n_i = n(N_i/N)$ 。如果第 i 层的标准差已知为 σ_i ，则第 i 层的抽样数目 n_i 为

$$n_i = n(N_i\sigma_i / \sum N_i\sigma_i)$$

(4) 在每一层，利用随机抽样抽出 n_i 个样本点。

(5) 如果要估计总体平均数，计算每一层的样本平均数 \bar{x}_i ，所以总体平均数的估计值 \bar{x} ，

$$\bar{x} = \left(\frac{N_1}{N}\right)\bar{x}_1 + \left(\frac{N_2}{N}\right)\bar{x}_2 + \cdots + \left(\frac{N_k}{N}\right)\bar{x}_k$$

(6) 计算每一层的样本方差 S_i^2 ，估计量 \bar{X} 的标准差 $S_{\bar{X}}$

$$S_{\bar{X}} = \sqrt{\left(\frac{N_1 - n_1}{N_1 - 1}\right)\left(\frac{N_1}{N}\right)^2 \frac{S_1^2}{n_1} + \left(\frac{N_2 - n_2}{N_2 - 1}\right)\left(\frac{N_2}{N}\right)^2 \frac{S_2^2}{n_2} + \cdots + \left(\frac{N_k - n_k}{N_k - 1}\right)\left(\frac{N_k}{N}\right)^2 \frac{S_k^2}{n_k}}$$

8.6 整群抽样

如果总体有许多组别，称为群（cluster），而每组中包含的个体性质差不多，每个丛集可以代表全部总体。例如：产品每 12 件打成一包为一组。整群抽样（cluster sampling）是以随机抽样抽出一组或多组。然后将这一组或多组的所有个体全部或部分，当作样本数据。整群抽样的进行步骤如下。

(1) 总体中的个体分成 N 组（丛集）。

(2) 决定抽样的组数 m ；如果抽出的丛集不要普查，则决定每个抽样出来的组中要再抽样的个数 k 。

(3) 以组为原始单位，进行随机抽样抽出 m 个组。

(4) 在每一抽样出来的组中，利用随机抽样抽出 k 个样本点。

(5) 如果要估计总体平均数，计算每一组的样本平均数 \bar{x}_i ，所以总体平均数的估计量 \bar{x}

$$\bar{x} = \frac{\vec{x}_1 + \vec{x}_2 + \cdots + \vec{x}_m}{m}$$

(6) 估计量 \bar{X} 的标准差 $S_{\bar{X}}$

$$S_{\bar{X}} = \sqrt{\frac{(N - m) \sum (\bar{x}_i - \bar{x})^2}{Nm(m - 1)}}$$

分层抽样（如图 8-4 所示）是层与层之间是异质个体（即不同性质分类的个体），而在同一层之内是同质个体。整群抽样（如图 8-5 所示）是从丛与丛之间是同质个体（即没有分类的个体），而在同一层之内可能存在异质个体。

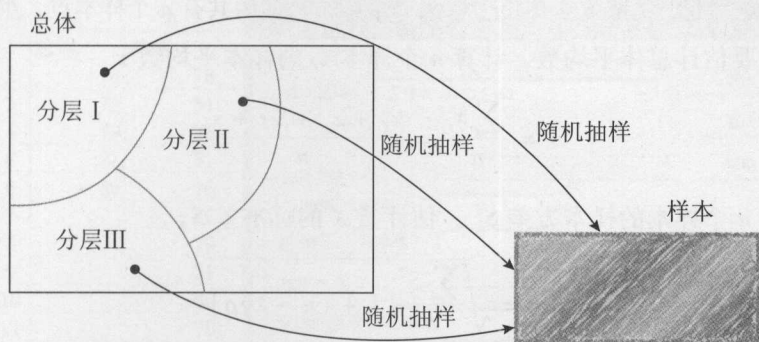


图 8-4 分层抽象

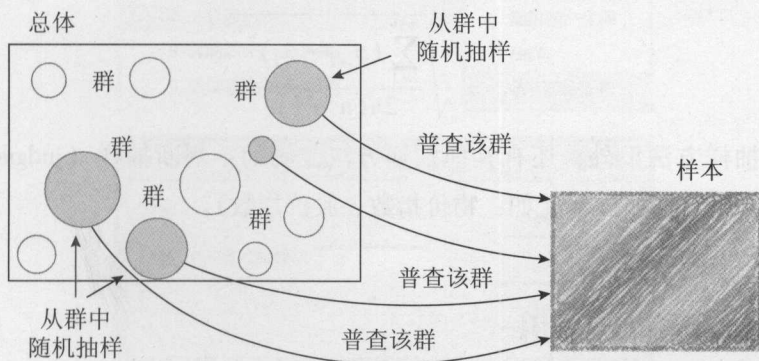


图 8-5 整群抽样

盖洛普（Gallup）民意调查，在美国大选年的抽样是利用多阶段的整群抽样。首先将美国分成东西南北四区（regions）（分层），每区利用简单随机抽样抽出几个城市（towns），每个城市分成数区（wards），利用简单随机抽样抽出几区，每区分成数里（precincts），利用简单随机抽样抽出几里，最后利用简单随机抽样抽出几个家庭，再进行访问调查。

8.7 系统抽样

系统抽样（system sampling）是将总体中的个体依序编号或排列，然后按照一个固定间隔，每隔若干号抽出一个样本点，组成一个样本。系统抽样的进行步骤如下。

- (1) 将总体中的个体随机排列。总体的容量为 N 。
- (2) 决定抽样的个数 n 。将已排列的总体，分成 n 个间隔。 $k = N/n$ 是间隔的长度。
- (3) 在第一个间隔，利用随机抽样抽出一个样本点，当作起点。
- (4) 从起点算起，每隔 k 个单位，取出一个样本，最后共有 n 个样本点，组成一个样本。
- (5) 如果要估计总体平均数，计算 n 个样本 x_i 的样本平均数 \bar{x}

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- (6) 计算 n 个样本的样本方差 S_i^2 ，估计量 \bar{X} 的标准差 $S_{\bar{X}}$

$$S_{\bar{X}} = \sqrt{\frac{\sum S_i^2}{n} [1 + (n-1)\rho]}$$

总体每对数据的相关系数为 ρ ，当总体是随机排列，即 $\rho = 0$ 。但是 ρ 是未知，上述公式无法计算。耶茨氏 (Yates) 建议下列公式

$$S_{\bar{X}} = \sqrt{\frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{2n(n-1)}}$$

除了以上抽样方法以外，还有其他抽样方法，例如：判断抽样 (judgment sampling) 是利用专家的判断，选择样本 (如：物价指数、股价指数)。

8.8 中文统计应用

8.8.1 抽样 (例题 8.1)

选择“中文统计”→“抽样理论”→“抽样”，在弹出的“抽样”进行下列操作，如图 8-6 所示。

- (1) 在“输入区域”下拉列表框中选取总体数据范围。
- (2) 如果输入区域有变量名称，则勾选“标记”。
- (3) “抽样方法”选择“随机”，并输入样本量“10”。
- (4) 选择输出选项，此范例选择“输出区域”，位置为 (B2)。单击“确定”。

8.8.2 中心极限定理 (例题 8.3)

执行“中心极限定理”的操作示意图和结果如图 8-7 所示。

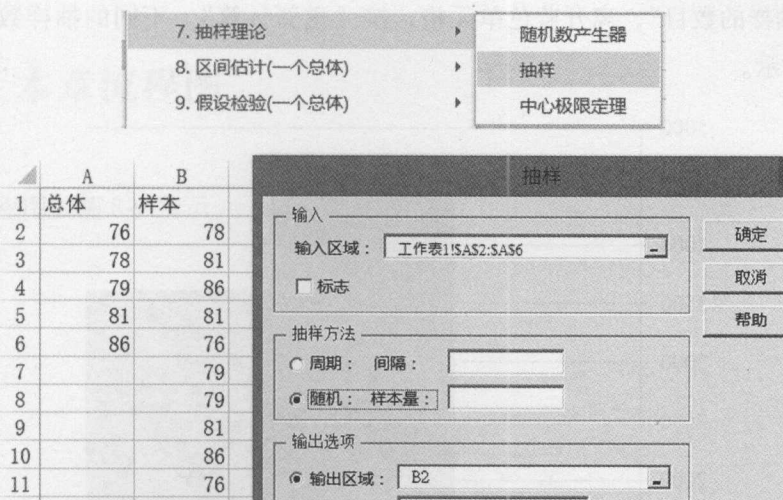


图 8-6 执行“抽样”的操作示意图

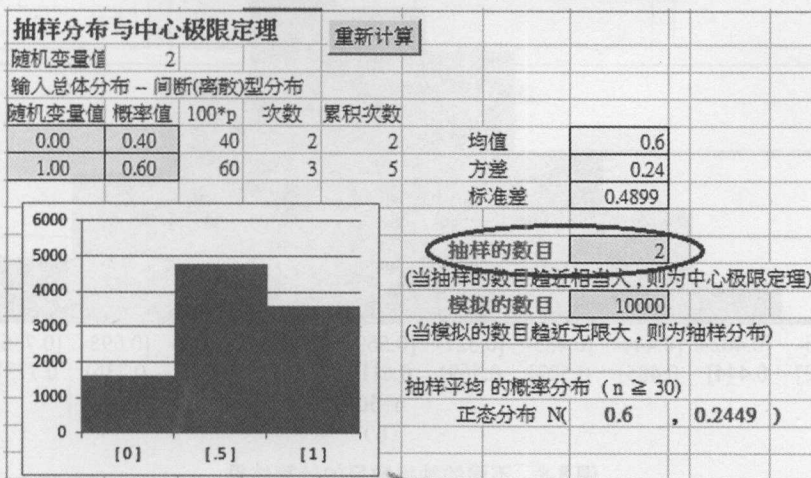
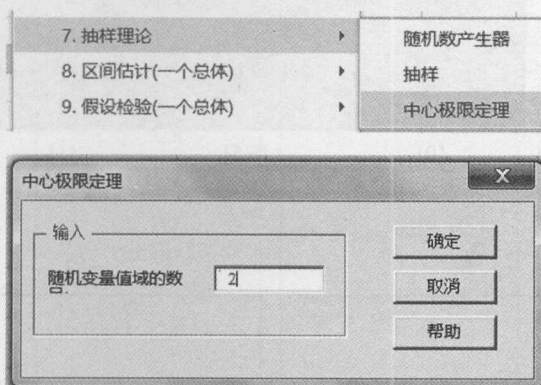


图 8-7 执行“中心级限定理”的操作示意图和结果

修改“抽样的数目”，离开蓝色单元格，按“重新计算”。不同的抽样数目的计算结果如图 8-8 所示。

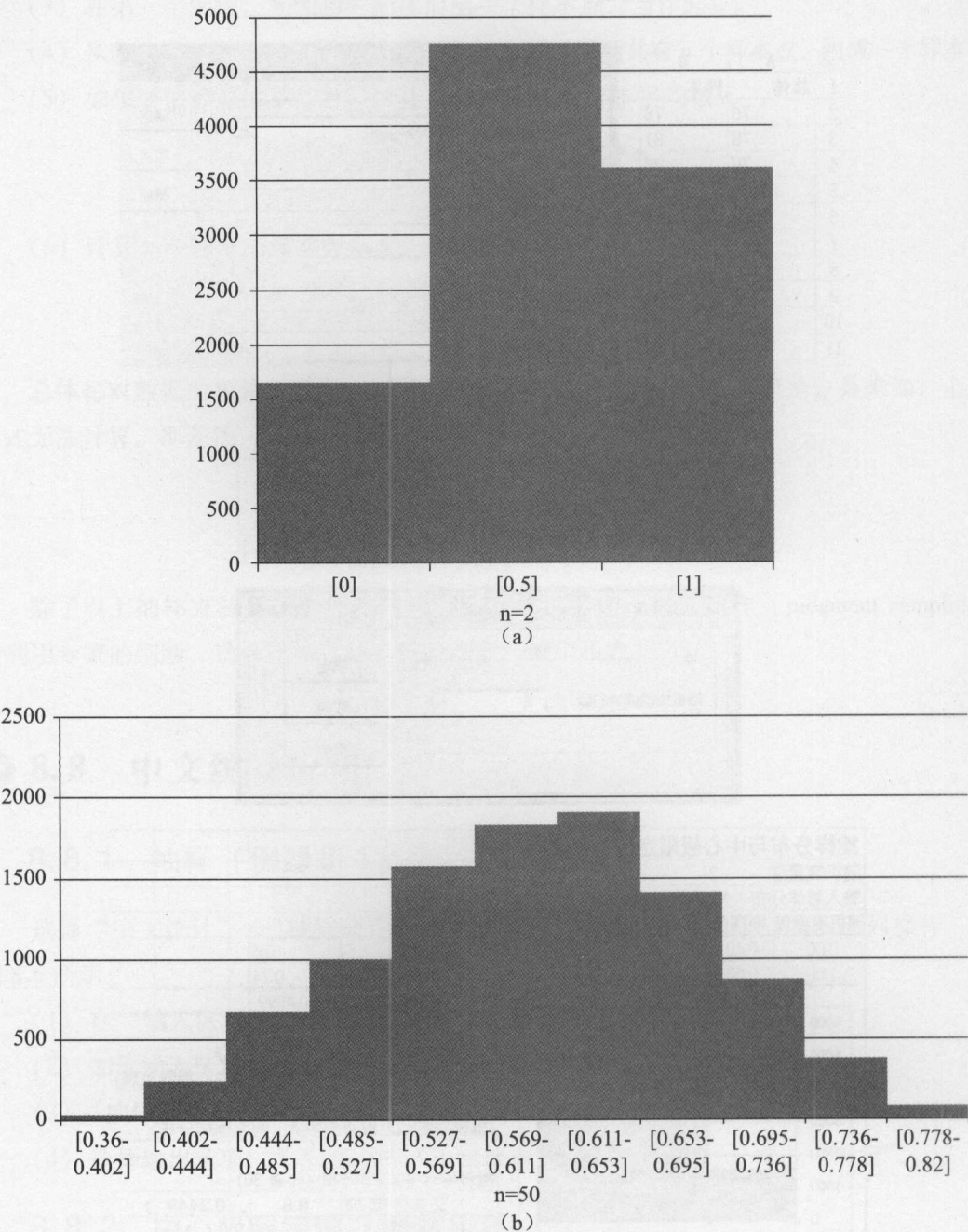


图 8-8 不同的抽样数目的计算结果

(a) $n=2$; (b) $n=50$

8.9 本章流程图

本章流程图如图 8-9 所示。

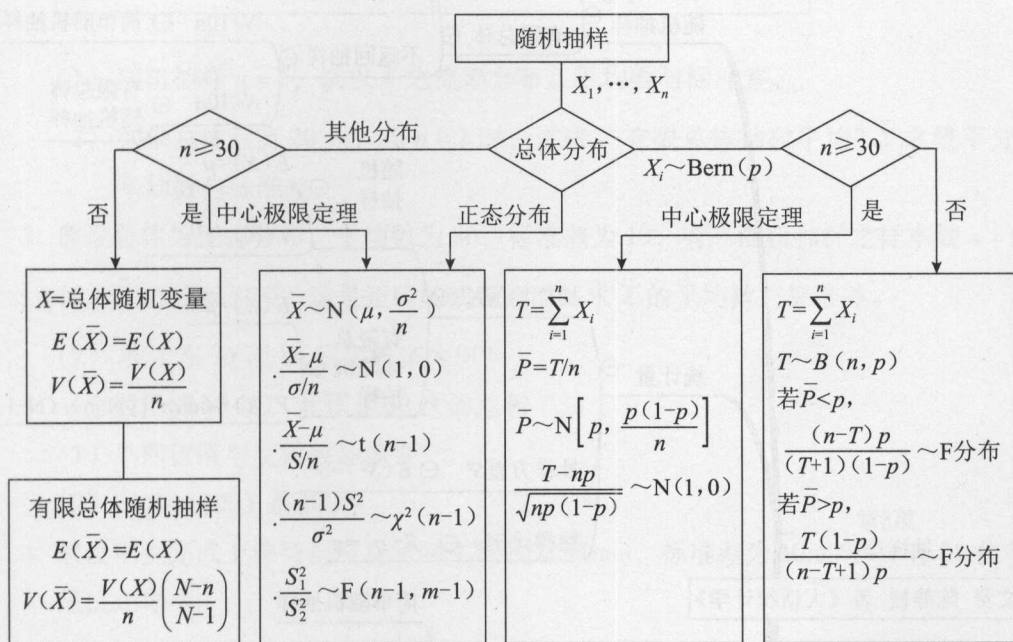


图 8-9 第 8 章流程图

8.10 本章思维导图

本章思维导图如图 8-10 所示。

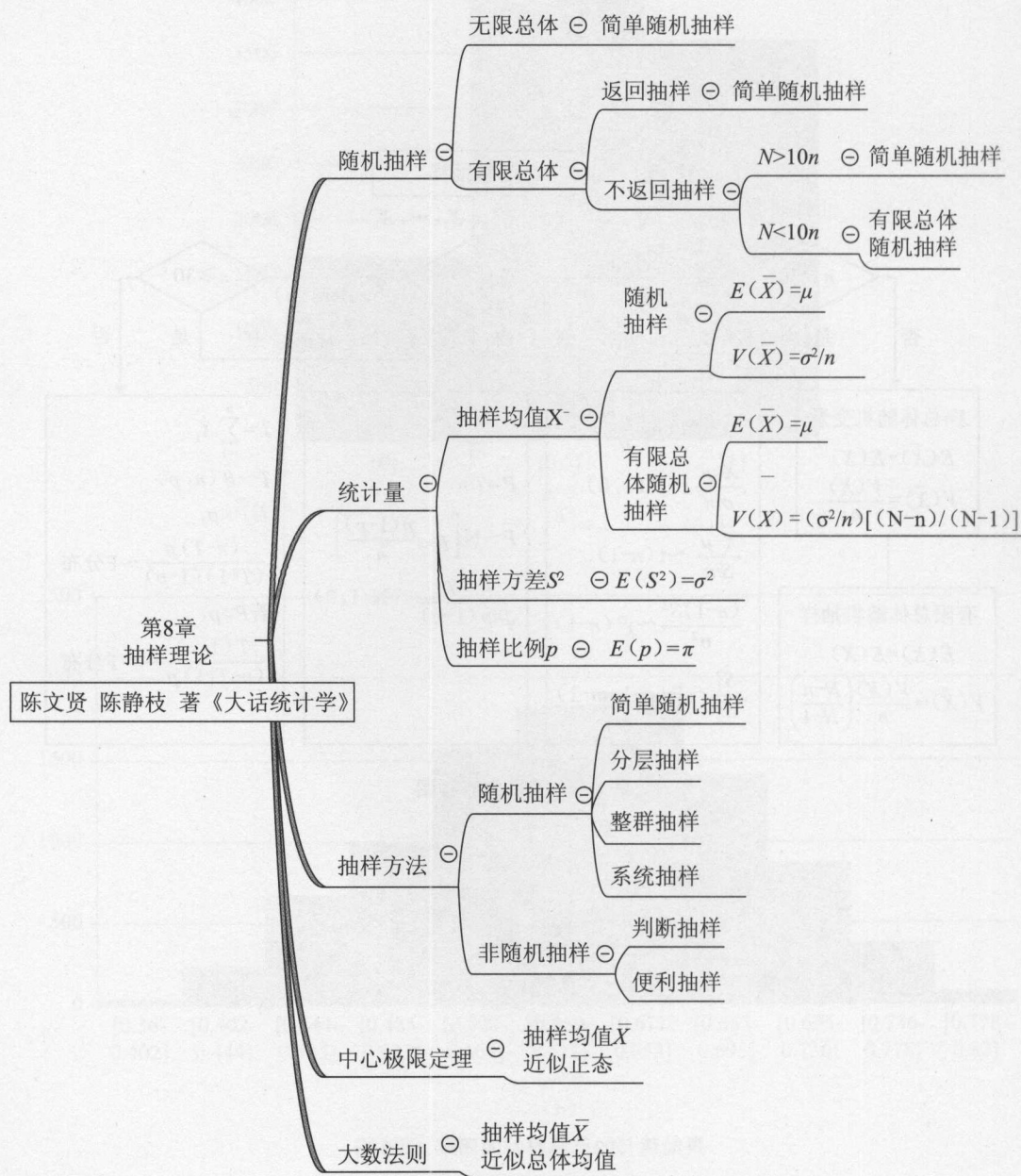


图 8-10 第 8 章思维导图

习 题

1. 假设某大学某系学生每月看电影次数呈下列之概率分布：

数目	概率 $P(x)$
0	0.5
1	0.4
2	0.1

- (1) 随机抽样 $n=2$ ，试求 \bar{X} 之概率分布、平均数与标准差。
- (2) 如果总体只有 20 人，当 $n=2$ 时，试求“有限总体抽样平均” \bar{X} 之概率分布、平均数与标准差。
2. 假设总体为正态分布，平均数为 80，标准差为 10，有一随机抽样之样本数 $n=9$ ：
 - (1) \bar{X} 之分布为何？这是正确的或近似的？求 \bar{X} 的平均数与标准差。
 - (2) 求 \bar{X} 在 76 至 84 间之概率为何？
3. 一个骰子掷 60 次，计算其 60 次的总和 T 。
 - (1) T 期望值与变数是多少？
 - (2) $P(T \leq 200)$ 是多少？
4. 假设学生完成上课登记所花时间平均数为 94min，标准差为 10min，今有 81 位学生随机之样本。
 - (1) 求 \bar{X} 的平均数与标准差。
 - (2) \bar{X} 之分配有何特性？
 - (3) $P(\bar{X} > 96)$ 。
 - (4) $P(92.3 < \bar{X} < 96)$ 。
 - (5) $P(\bar{X} < 95)$ 。
5. 假设某堆行李平均质量为 55 磅 (24.9476kg)，标准差为 7 磅 (3.1751kg)。如果随机抽取 40 件行李为样本，样本平均质量 \bar{X} 位于 54 磅 (24.4940kg) 至 56 磅 (25.4012kg) 间之概率为何？

其他习题请下载。



第9章

统计估计

欲观千岁，则数今日；欲知亿万，则审一二。

——荀子《非相篇》

子乃规规然而求之以察，索之以辩，是直用管窥天，用锥指地也，不亦小乎？

——《庄子·秋水篇》

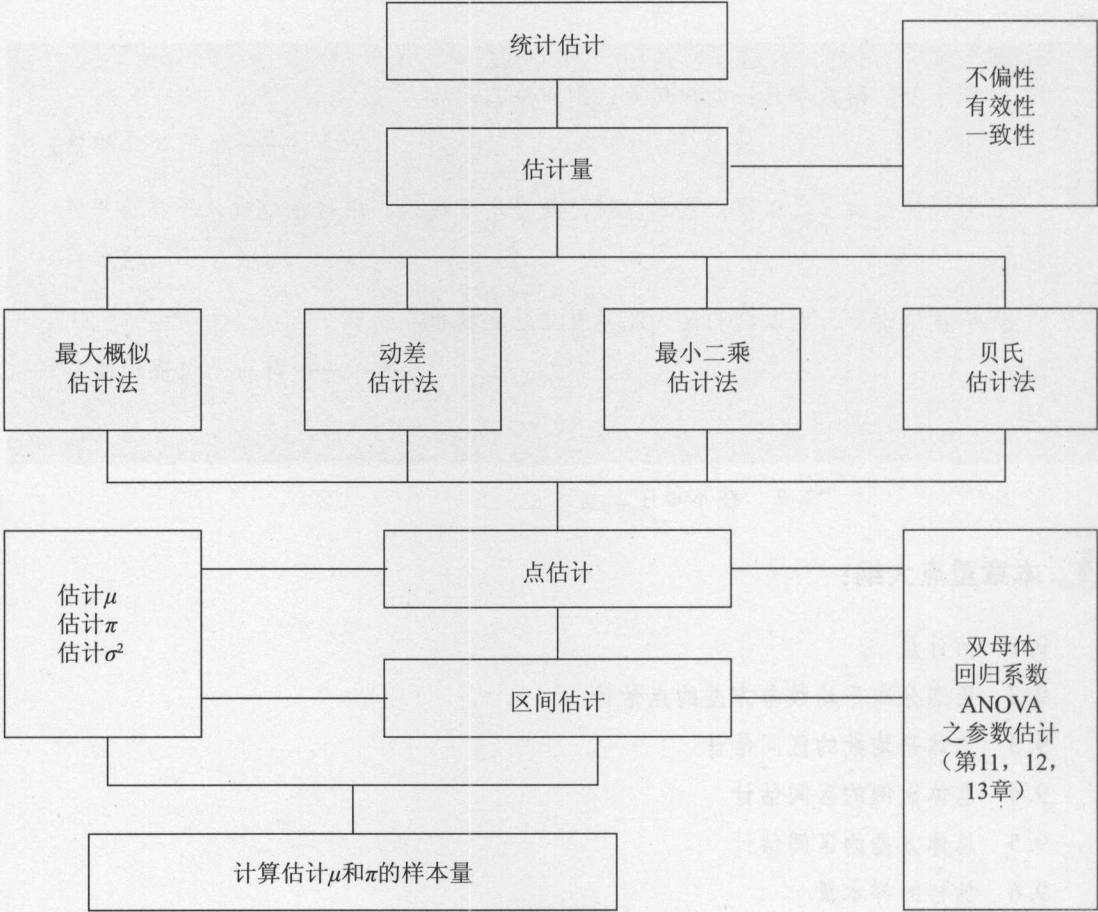
不遇盘根错节，无以别利器，此乃吾立功之秋也。

——司马光《资治通鉴》



本章重点大纲：

- 9.1 估计量
- 9.2 正态分布平均数与方差的点估计
- 9.3 总体平均数的区间估计
- 9.4 总体比例的区间估计
- 9.5 总体方差的区间估计
- 9.6 抽样的样本量
- 9.7 标准误差
- 9.8 中文统计应用
- 9.9 本章流程图
- 9.10 本章思维导图



本章概念图

9.1 估计量

统计估计有：点估计 (point estimate) 和区间估计 (interval estimate)。两者都需要估计量。

定义 假设 X_1, X_2, \dots, X_n 为 n 个抽样随机变量，总体参数的一个估计量 (estimator)，是这 n 个抽样随机变量的函数 $f(X_1, X_2, \dots, X_n)$ ，但这个函数中不包括未知参数。

估计量不是唯一的，一个参数会有很多估计量。

例如： $\bar{X} = \frac{(X_1 + X_2 + \dots + X_n)}{n}$ 是总体平均数 μ 的一个估计量。

例如： σ 未知， $\frac{(X_1 + X_2 + \dots + X_n + \sigma)}{n}$ 不是总体平均数 μ 的估计量。

例如： μ 未知， $\frac{\sum (X_i - \mu)^2}{n}$ 不是总体方差 σ^2 的估计量。但 $\frac{\sum (X_i - \bar{X})^2}{n-1}$ 是 σ^2 的估计量。

估计量是一个统计量 (随机变量)，只是估计量有其估计对象，即某个总体参数。通常将参数 θ 的估计量记作 $\hat{\theta}$ 。在本章中，有时也用 U, V, W 代表同一参数的不同估计量。

定义 假设 x_1, x_2, \dots, x_n 为 n 个抽样值，一个总体参数的估计量值，简称估计值 (estimate)，是这 n 个样本随机值的函数值 $f(x_1, x_2, \dots, x_n)$ ，估计值即为点估计。

例如： $\bar{x} = \frac{(x_1 + x_2 + \dots + x_n)}{n}$ 是一个估计值。

估计量是随机变量，应该有其抽样分布、期望值与方差，主要是理论统计的证明。估计值是一个实数值，因不同的抽样结果，而有不同的值，主要是应用统计的计算结果。

总体参数的估计量不是唯一的，如何选择其估计量，就要考虑估计量的无偏性、有效性与一致性。

定义 一个总体参数的估计量的期望值，与参数值之间的差距，称为该估计量的偏误 (bias)。如果一个估计量的期望值，等于参数值，即其偏误为零，则称该估计量为无偏估计量 (unbiased estimator)。

定义 一个估计量 U 的均方误 (mean squared error)，记作 $MSE(U)$

$$\begin{aligned} MSE(U) &= E(\text{估计量} - \text{参数})^2 = \text{估计量减参数的平方的期望值} \\ &= (\text{估计量的方差}) + (\text{估计量的偏误})^2 \end{aligned}$$

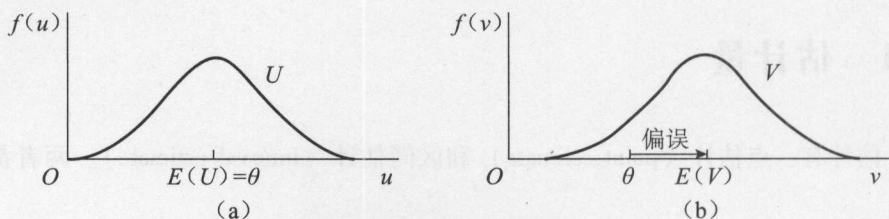


图 9-1 有偏估计量与无偏估计量

(a) 无偏估计量; (b) 有偏估计量

通常利用 MSE 来比较两个估计量 U, W 的效率 (efficiency)。MSE 越小, 效率越大。

(有的教科书将效率定义为 $\frac{V(W)}{V(U)}$, 只考虑无偏估计量的方差, 没有考虑偏误)。

若估计量 U 比估计量 V 的效率 $\frac{\text{MSE}(W)}{\text{MSE}(U)} > 1$, 则估计量 U 比估计量 W 好 (有效)。

偏误大小称为“准度”, 方差大小称为“精度”, 合起来称为“精准度”。图 9-2 是估计量的有效性之比较。

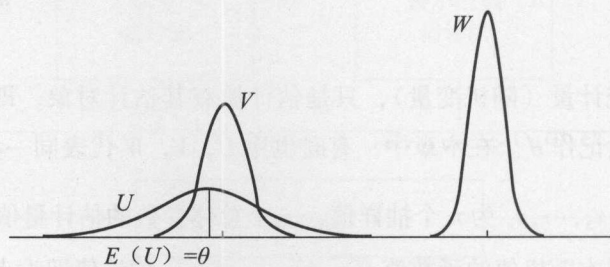


图 9-2 估计量的有效性之比较

如图 9-2 所示, θ 是目标参数。 U 是不偏估计量, 方差最大, 准度高。 V 有一点偏误, 方差较小。 W 的偏误最大, 方差最小, 精度高。

如果一个估计量, 当样本量越大, 它的概率越集中在估计参数上, 称为一致性。换言之, 当样本量越大, 估计值和参数的误差越小。其数学定义如下。

定义 若 $\hat{\theta}$ 为 θ 的估计量, 对任何实数 $\delta > 0$, $\lim_{n \rightarrow \infty} P\{|\hat{\theta} - \theta| < \delta\} = 1$, 则 $\hat{\theta}$ 为 θ 的一致性估计量 (consistent estimator)。

定义 若 $\hat{\theta}$ 为 θ 的估计量, $\lim_{n \rightarrow \infty} E(\hat{\theta}) = \theta$, 则 $\hat{\theta}$ 为 θ 的近似无偏估计量 (asymptotically unbiased estimator)。

定理 若 $\hat{\theta}$ 为 θ 的一致性估计量, 则 $\hat{\theta}$ 为 θ 的近似无偏估计量。

定理 若 $\hat{\theta}$ 为 θ 的一个估计量, $\lim_{n \rightarrow \infty} \text{MSE}(\hat{\theta}) = 0$, 则 $\hat{\theta}$ 为 θ 的一致性估计量。

一致性估计量, 是当抽样数目(样本量)越多, 则估计量和参数的误差越小。

参数的点估计量的求法有下列方法: 最大似法、动差法、贝氏法、最小二乘法。

9.2 正态分布平均数与方差的点估计

9.2.1 正态分布平均数的点估计

假设 X_1, X_2, \dots, X_n 为 n 个抽样随机变量, $X_i \sim N(\mu, \sigma^2)$ 。总体平均数 μ 的点估计, 其估计量通常用抽样平均 \bar{X}

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

定义 一个统计量的标准差, 称为该统计量的标准误差 (estimated standard error)。

抽样平均 \bar{X} 的标准误差是 σ/\sqrt{n} 。

若 X_i 的标准差 (即总体标准差) σ 未知, 则估计量 \bar{X} 的标准差还是 σ/\sqrt{n} , 但是不能算出, 要代入抽样标准差 s 。抽样平均 \bar{X} 的标准误差是 s/\sqrt{n} 。

9.2.2 正态分布方差的点估计

假设 X_1, X_2, \dots, X_n 为正态分布的抽样随机变量, 总体方差的点估计是估计量 S^2

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

总体标准差的点估计是估计量 S

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

估计量 S^2 是方差 σ^2 的无偏估计量。

抽样方差 S^2 的标准差是 $\sqrt{2}\sigma^2/\sqrt{n-1}$ 。但是, 很少称抽样方差 S^2 的标准误差。

9.3 总体平均数的区间估计

区间估计, 是利用抽样数据, 决定一个区间, 该区间为某参数可能落入的范围, 即以区间形式估计参数值的大小。

定义 假设 X_1, X_2, \dots, X_n 为 n 个抽样随机变量。令 $A = f(X_1, X_2, \dots, X_n)$ 为一个抽样估计量, $B = g(X_1, X_2, \dots, X_n)$ 为另一个抽样估计量。若 $P(A \leq \theta \leq B) = 1 - \alpha$, θ 为总体参数, 则称 $[A, B]$ 为 θ 的 $1 - \alpha$ 置信区间 (confidence interval)。 $1 - \alpha$ 称为置信度或置信系数 (confidence coefficient)。 A 为置信下限, B 为置信上限。

实际上, 置信区间是将随机抽样值 x_i 代入 $[A, B]$ 。所以, 参数是常数, 是固定的, 但是未知的。置信区间是变动的, 不同的抽样结果, 产生不同的置信区间。

通常置信区间是估计值 (样本统计值) 两边各加一个常数。方差 σ^2 的置信区间, 并非估计值 s^2 加减一个半径 E , 而是除以两个常数。

如果 σ 未知, 则以样本标准差 s 代替 σ , 这样会降低置信区间的置信度。所以为了保持置信度, 要放宽置信区间, 于是用较大的 t 值 $t_{\alpha/2}$ 代替 $z_{\alpha/2}$ 。平均数的区间估计如表 9-1 所示。

- 在固定的置信度之下, 置信区间的长度越小越好, 表示估计较可信, 如图 9-3 所示。
- (1) 当样本容量 n 越大, 置信区间的长度 L 就越小。
 - (2) 当总体标准差 σ 越大, 置信区间的长度 L 就越大。
 - (3) 当置信度 $1 - \alpha$ 越大, 置信区间 L 的长度就越大。

表 9-1 平均数的区间估计

条件一	条件二	条件三	置信区间
X_i 为正态分布	σ 已知	有限总体 $N > 20n$ 或无限总体	$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$ (如图 9-5)
		有限总体 $N < 20n$	$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right]$ (如图 9-6)
	σ 未知	有限总体 $N > 20n$ 或无限总体	$\left[\bar{x} - t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}} \right]$
		有限总体 $N < 20n$	$\left[\bar{x} - t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \bar{x} + t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right]$
X_i 为非正态分布	σ 已知	$n < 30$	不能利用 Z 或 t 分布做区间估计, 可利用非参数统计
		$n > 30$	$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$
		$n < 30$	不能用 Z 或 t 分布做区间估计, 可利用非参数统计
	σ 未知	$30 < n < 100$	$\left[\bar{x} - t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}} \right]$
		$n \geq 100$	$\left[\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right]$

置信区间长度 L , 样本容量 n , 置信度 $1 - \alpha$, 标准差 $\sigma(s)$, 4 个变量的关系, 如果其中两个固定, 另外两个变量正负相关, 如图 9-4 所示。负相关是一个变量变大, 另一个变量变小。

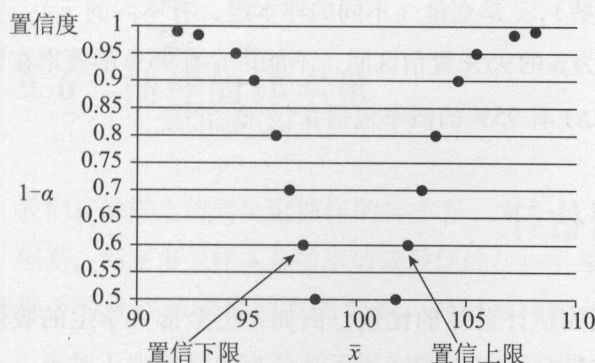


图 9-3 置信度与置信区间长度的关系

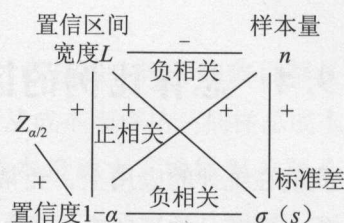


图 9-4 $L, n, 1 - \alpha, \sigma(s)$ 两两关系图

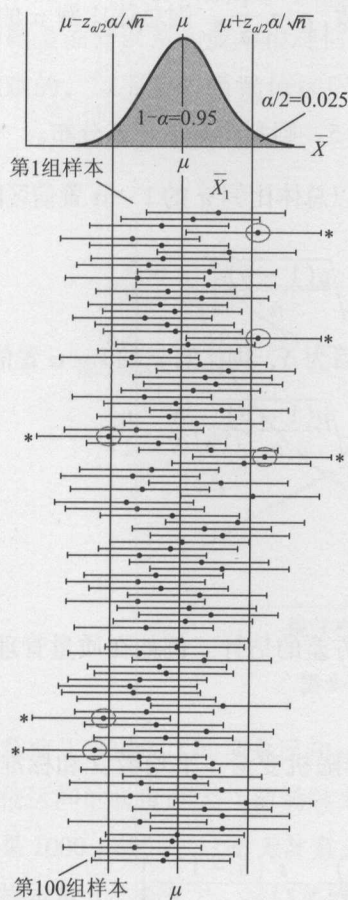


图 9-5 总体方差已知

- 注：(1) 以相同的样本数计算 100 个 95% 置信区间，有 6 个置信区间不含总体均值。(有 *)
- (2) 每个置信区间的长度相同。
- (3) 每组样本有 n 个数据。

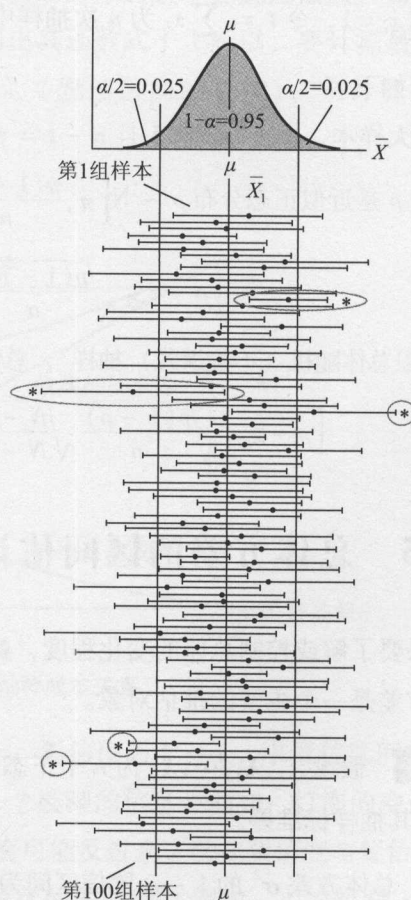


图 9-6 总体方差未知

- 注：(1) 以相同的样本数计算 100 个 95% 置信区间，有 5 个置信区间不含总体均值。(有 *)
- (2) 每个置信区间的长度不同。
- (3) 每组样本有 n 个数据。

因为参数 μ 是未知常数 (不变的数), \bar{x} 是变量 (不同的样本组, 有不同的 \bar{x}), 置信区间 $[A, B]$ 是变动的。所以, $[A, B]$ 为 μ 的 95% 置信区间, 不能说 μ 有 95% 的概率在置信区间 $[A, B]$, 而应该说置信区间 $[A, B]$ 有 95% 的概率包括 μ 。

9.4 总体比例的区间估计

我们在统计的应用上, 经常要预先估计总体的比例, 例如: ①全部大学生的吸烟比例; ②一批产品的不良率; ③全台北市的大选投票率。

假设 X_1, X_2, \dots, X_n 为 n 个贝努里分布的抽样随机变量, x_i 为其随机值, $x_i = 0$ 代表第 i 次抽样或 $x_i = 1$, 令 $t = \sum_{i=1}^n x_i$ 为 n 次抽样中成功的次数, $p = \frac{t}{n}$, 则总体比例 π 的区间估计, 计算如下。

n 为大样本, 即 $t = np \geq 5$ 且 $n - t = n(1 - p) \geq 5$, 则利用标准正态分布。

因为 p 是近似正态分布 $p \sim N\left[\pi, \frac{\pi(1-\pi)}{n}\right]$, 所以总体比例 π 的 $1 - \alpha$ 置信区间为

$$\left[p - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right]$$

“有限总体随机 (不投返式) 抽样”, 总体的总体数目为 N , 则比例 π 的 $1 - \alpha$ 置信区间为

$$\left[p - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n} \frac{N-n}{N-1}}, p + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n} \frac{N-n}{N-1}} \right]$$

9.5 总体方差的区间估计

如果要了解或控制总体的变化程度, 就需要总体方差的估计。例如在质量管理的控制图中, 方差是一个主要的推估对象。

定理 假设 X_1, X_2, \dots, X_n 为 n 个正态分布的抽样随机变量, 平均数 μ 和标准差 σ 未知, s 为其抽样标准差。

(1) 总体方差 σ^2 的 $1 - \alpha$ 置信区间为 $\left[\frac{s^2(n-1)}{\chi_{\alpha/2}^2(n-1)}, \frac{s^2(n-1)}{\chi_{1-\alpha/2}^2(n-1)} \right]$ 。

(2) 总体标准差 σ 的 $1 - \alpha$ 置信区间为 $\left[\sqrt{\frac{s^2(n-1)}{\chi_{\alpha/2}^2(n-1)}}, \sqrt{\frac{s^2(n-1)}{\chi_{1-\alpha/2}^2(n-1)}} \right]$ 。

例题 9.1—9.4 (见网络资源)

9.6 抽样的样本量

我们在抽样之前要决定抽样的样本量，对于统计估计和检验，样本的数目都是已知数。但是，到底多少样本是适当的或最佳的？一个考虑是从成本来评估，抽样总成本包括抽样误差的成本和收集样本的成本。根据抽样理论，样本量越大，抽样误差越小（ σ/\sqrt{n} ），再加上实际抽样多数是不返回抽样 [有限抽样因子 $(N-n)/(N-1)$]，所以误差会更小。样本量越大，收集抽样成本越大。因此，会有一个最低的抽样总成本，其对应的样本量即为最佳样本量，如图 9-7 所示。收集抽样的成本是直线函数，而且容易估计，问题是抽样误差的成本很难估计，更难列出其函数式子。所以，要计算最佳样本量是有困难的，以下我们用置信区间的信息意义（置信区间的长度），来计算抽样的样本量。

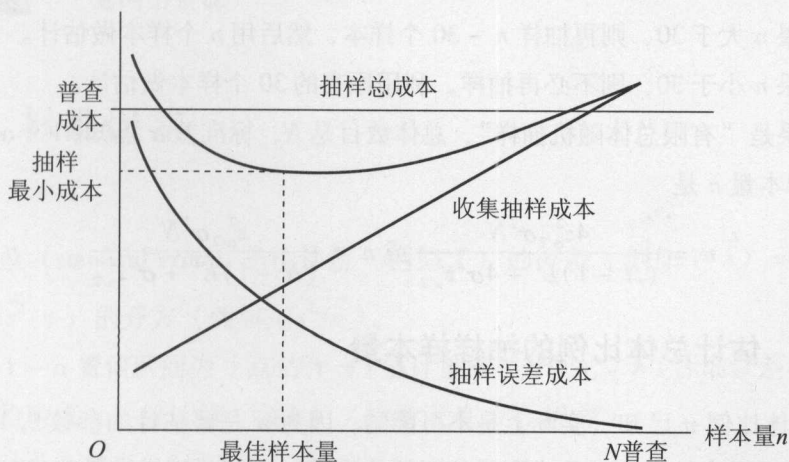


图 9-7 样本量与抽样成本关系

本章前几节置信区间都是已知：样本量为 n ，置信度为 $1 - \alpha$ ，求置信区间的长度 L 。但是置信区间的长度越大（范畴越大），越是一个模糊的信息。例如：灯泡的寿命 95% 置信区间是 $1000 \pm 600\text{h}$ ，这是无甚意义的。现在有可能反过来，已知置信度与置信区间的长度，求样本量 n 。换言之，我们要固定一个置信区间的长度（例如：平均数在 ± 20 之内，比例在 $\pm 1\%$ 之间），于是在某个置信水平（例如：95%）之下，应该抽样多少。

9.6.1 估计总体平均数的抽样样本量

1) 如果总体标准差 σ 已知, $1 - \alpha$ 置信区间的长度 L (区间全长, 有的公式取置信区间的半径, 置信区间为 $p \pm E$) 确定, 则样本量 n 是

$$n = \left[\frac{2z_{\alpha/2}\sigma}{L} \right]^2 \text{ 或 } n = \left[\frac{z_{\alpha/2}\sigma}{E} \right]^2$$

2) 如果总体标准差 σ 未知, $1 - \alpha$ 置信区间的长度 L (区间全长, 有的公式取置信区间的半径, 置信区间为 $p \pm E$) 确定, $L = 2E$, 以历史数据或过去经验估计标准差 s 。或者随机抽样 30 个样本, 计算样本标准差 s 。如果数据呈正态分布, 则可以用全矩除以 4 概略估计 s 。样本量 n 是

$$n = \left[\frac{2z_{\alpha/2}s}{L} \right]^2 \text{ 或 } n = \left[\frac{z_{\alpha/2}s}{E} \right]^2$$

上述公式按照精确计算, $z_{\alpha/2}$ 应该用 $t_{\alpha/2}$, 但是其自由度又牵涉到样本量, 而且 $n = 30$, $t_{\alpha/2}$ 近似 $z_{\alpha/2}$, 所以还是用 Z 分布。

(1) 如果 n 大于 30, 则再抽样 $n - 30$ 个样本, 然后用 n 个样本做估计。

(2) 如果 n 小于 30, 则不必再抽样, 利用原有的 30 个样本做估计。

(3) 如果是“有限总体随机抽样”, 总体数目是 N , 标准差 σ 已知, $1 - \alpha$ 置信区间的长度 L , 则样本量 n 是

$$n = \frac{4z_{\alpha/2}^2\sigma^2N}{(N-1)L^2 + 4\sigma^2z_{\alpha/2}^2} \text{ 或 } n = \frac{z_{\alpha/2}^2\sigma^2N}{(N-1)E^2 + \sigma^2z_{\alpha/2}^2}$$

9.6.2 估计总体比例的抽样样本量

1) 如总体比例 π 已知 (实际上是不可能的, 因为 π 是要估计的参数), $1 - \alpha$ 置信区间的长度 L (区间全长, 有的公式取置信区间的半径) 确定, 则样本量 n 是

$$n = \left[\frac{2z_{\alpha/2}}{L} \right]^2 p(1-p)$$

2) 如总体比例 π 未知, $1 - \alpha$ 置信区间的长度 L 确定, 则随机抽样 30 个样本做事前检验, 计算样本比例 p 。样本量 n 是

$$n = \left[\frac{2z_{\alpha/2}}{L} \right]^2 p(1-p)$$

(1) 如果 n 大于 30, 则再抽样 $n - 30$ 个样本。

(2) 如果 n 小于 30, 则不必再抽样, 利用原有的 30 个样本。

3) 如总体比例 π 未知, 而不想做事前检验, 取 $\pi(1-\pi)$ 的最大值 $1/4$ 。 $1 - \alpha$ 置信

区间的长度 L 确定, 则样本量 n 是

$$n = \left[\frac{z_{\alpha/2}}{L} \right]^2$$

4) 如果是“有限总体随机抽样”, 总体数目是 N , $1 - \alpha$ 置信区间的长度 L 确定, 则样本量 n 是

$$n = \frac{4z_{\alpha/2}^2 \hat{p}(1 - \hat{p})N}{(N - 1)L^2 + 4\hat{p}(\hat{p} - 1)z_{\alpha/2}^2}$$

以上结果告诉我们, 要检验总体比例 (例如: 总统直接民选, 某候选人的得票率), 要使预测的置信区间有 99% 的置信度, 置信区间的差距只有 5% (置信区间为 $p \pm 2.5\%$), 那么不管总体有多大 (1000 万人也好, 1 亿人也好), 我们只要抽样 2663 人。

盖洛普民意调查 (Gallup poll) 常用的样本量是 1823 人。

当然, 先决条件是, 抽样没有非抽样误差, 也就是样本的来源没有偏向某一特别阶层, 被抽样的受访者没有违心之答, 等等。

例题 9.5 (见网络资源)

9.7 标准误差

标准误差 (standard error) 是估计量 (例如 \bar{X}) 的方差 [例如 $V(\bar{X}) = \sigma^2/n$] 的估计值 (例如 s^2/n) 的开方 (例如 $\sqrt{s^2/n}$)。

参数的 $1 - \alpha$ 置信区间为 [点估计 \pm (估计量概率分配) $_{\alpha/2} \times$ (标准误差)]

参数的检验值 = (统计值 - 参数) / 标准误差 (第 10 章)

区间估计总表如表 9-2 所示。

表 9-2 区间估计总表

估计	参数	估计量	概率分配	标准误差
单总体平均数 (σ^2 已知)	μ	\bar{X}	Z	$\frac{\sigma}{\sqrt{n}}$
单总体平均数 (σ^2 未知)	μ	\bar{X}	$t(n-1)$	$\frac{s}{\sqrt{n}}$
单总体比例	π	p	Z	$\sqrt{\frac{p(1-p)}{n}}$

续表

估计	参数	估计量	概率分配	标准误差
单总体方差	σ^2	s^2	$\chi^2(n-1)$	\times
双总体平均数 (σ_1^2, σ_2^2 已知)	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	Z	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
双总体平均数 (σ_1^2, σ_2^2 未知, 相等)	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$t(n-2)$ $n = n_1 + n_2$	$s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
双总体平均数 (σ_1^2, σ_2^2 未知, 不等)	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$t(\nu)$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
双总体比例	$\pi_1 - \pi_2$	$p_1 - p_2$	Z	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
双总体方差	$\frac{\sigma_1^2}{\sigma_2^2}$	$\frac{s_1^2}{s_2^2}$	$F(n_1, n_2)$	\times
一因素 ANOVA	请见表 12-1			
二因素 ANOVA	请见表 12-2			
回归与相关	请见表 13-3			

标准差和标准误差的差别，标准误差是推断统计（估计和检验）的基础：

总体分布 X_i ，总体方差 $V(X_i) = \sigma^2$ ，总体标准差 σ

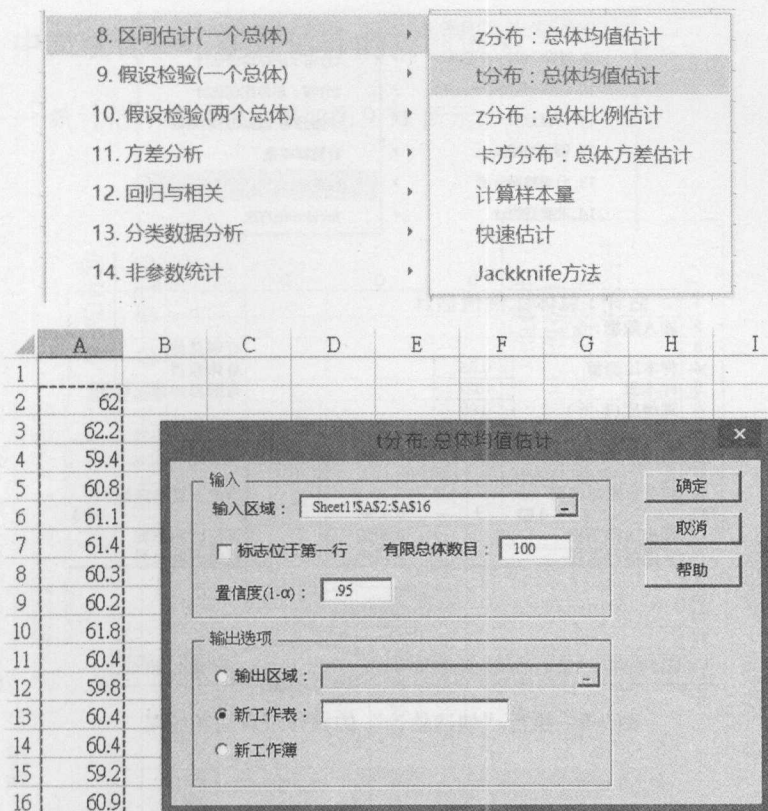
→ 总体均值 μ 的估计量（抽样平均） \bar{X} ，标准误差 $V(\bar{X}) = \sigma^2/n$ ，样本数据 x_i ，样本方差 $s^2 = \sum (x_i - \bar{x})^2/(n-1)$ ，样本标准差 s

→ 标准误差是估计量的标准差 s/\sqrt{n} （标准误差越小，推断结论越好）

9.8 中文统计应用

9.8.1 t 分布估计（例题 9.2）

执行“中文统计”→“区间估计（一个总体）”→“t 分布：总体均值估计”的操作示意图和结果如图 9-8 所示。



	A	B	C	D	E	F	G
1	t 分布: 总体均值估计						
2							
3			样本 1				
4	置信度(1-α)		0.95				
5	样本量		15				
6	样本均值		60.68667				
7	样本标准差		0.89751				
8					有限总体的数目		100
9	置信区间下界		60.18964		置信区间下界		60.22612
10	置信区间上界		61.18369		置信区间上界		61.14721

图 9-8 执行“t 分布: 总体均值估计”的操作示意图和结果

9.8.2 快速估计 (例题 9.3)

执行“快速估计”的操作示意图和结果如图 9-9 所示。

9.8.3 计算样本量 (例题 9.5)

执行“计算样本量”的操作示意图和结果如图 9-10 所示。



图 9-9 执行“快速估计”的操作示意图和结果

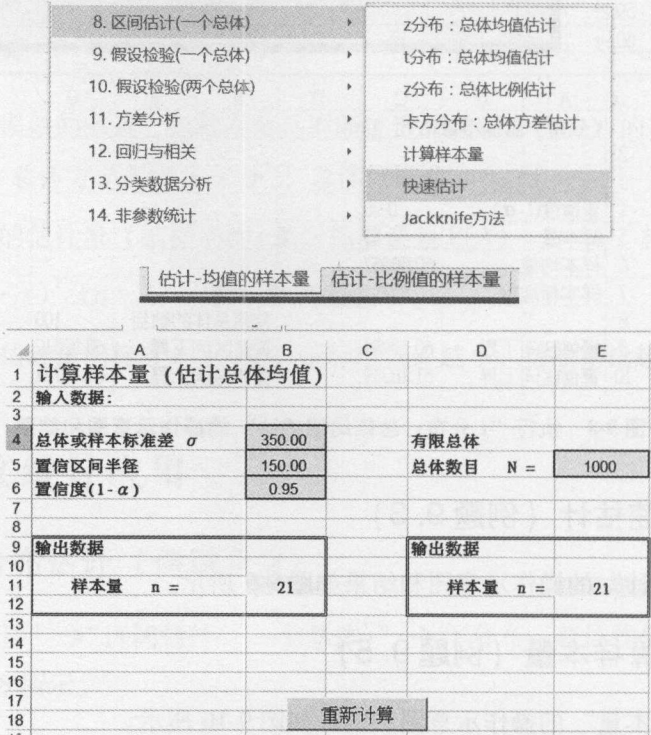


图 9-10 执行“计算样本量”的操作示意图和结果

9.8.4 中文统计——统计估计的功能地图

中文统计——统计估计的功能地图如图 9-11 所示。

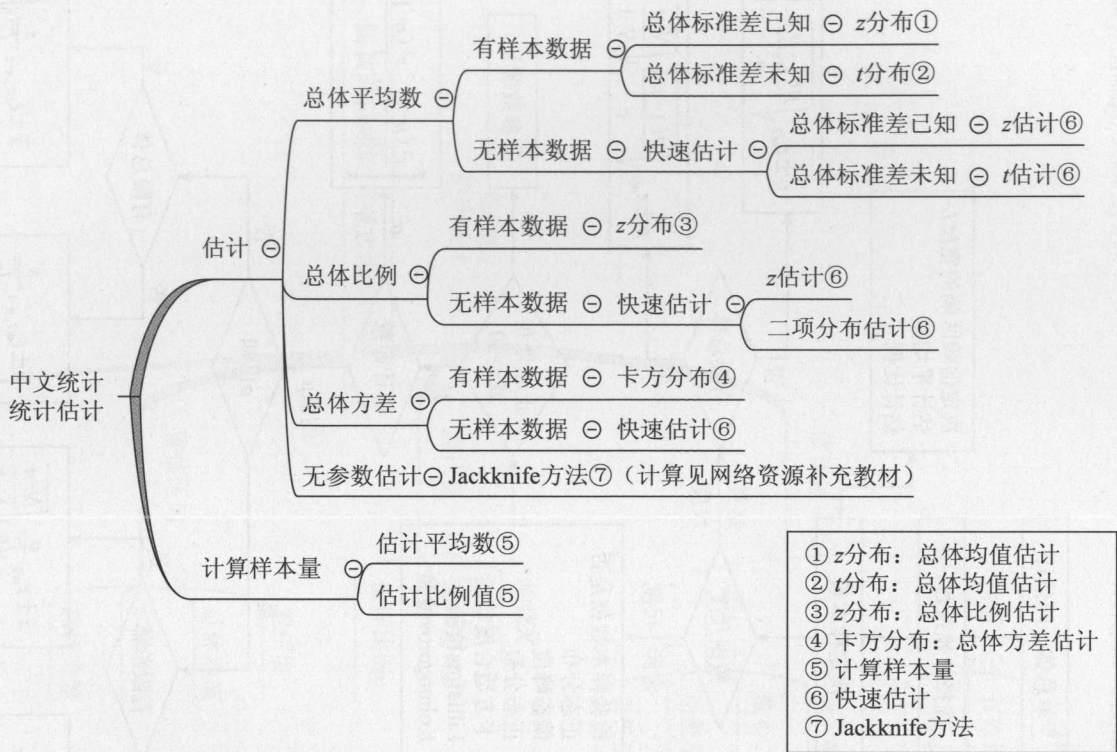


图9-11 中文统计——统计估计的功能地图

9.9 本章流程图

本章流程图如图 9-12 所示。

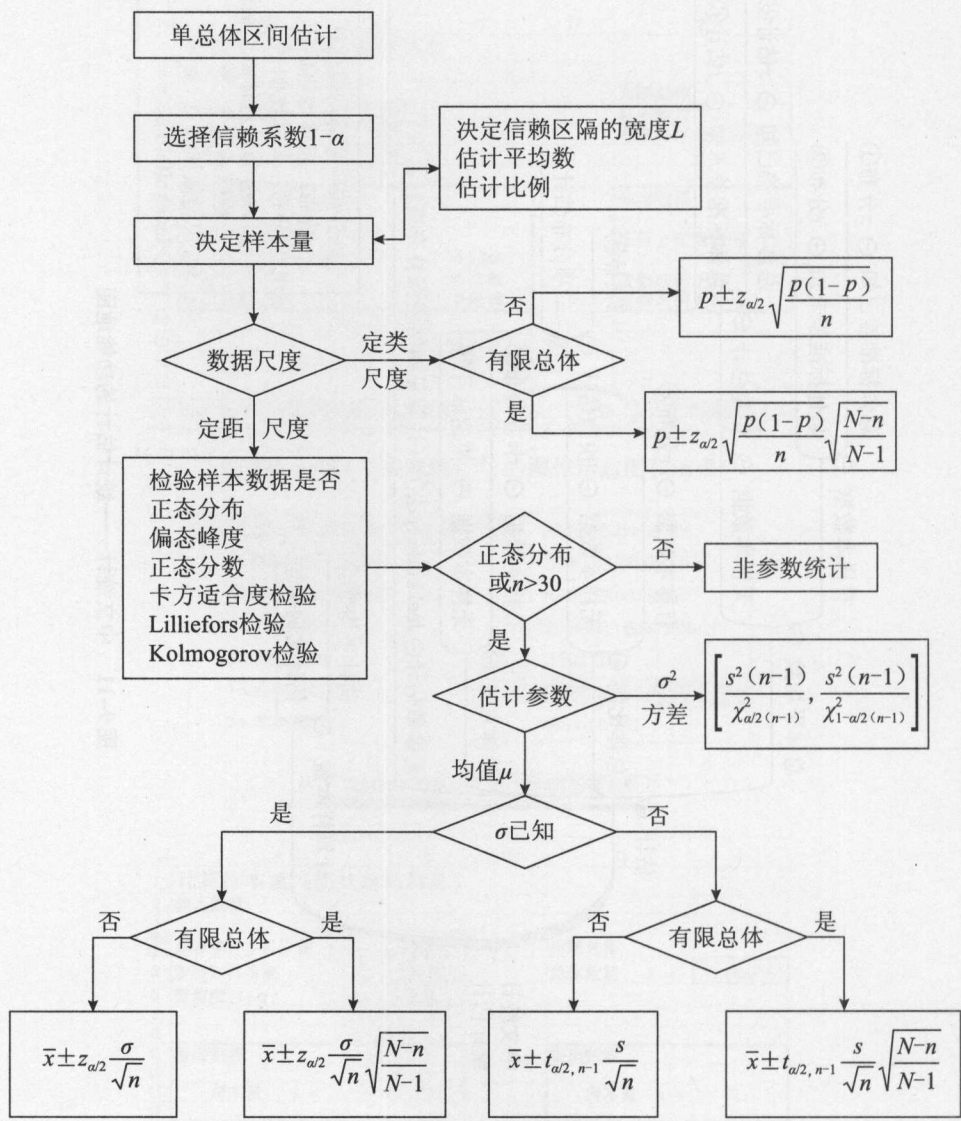


图 9-12 第 9 章流程图

9.10 本章思维导图

本章思维导图如图 9-13 所示。

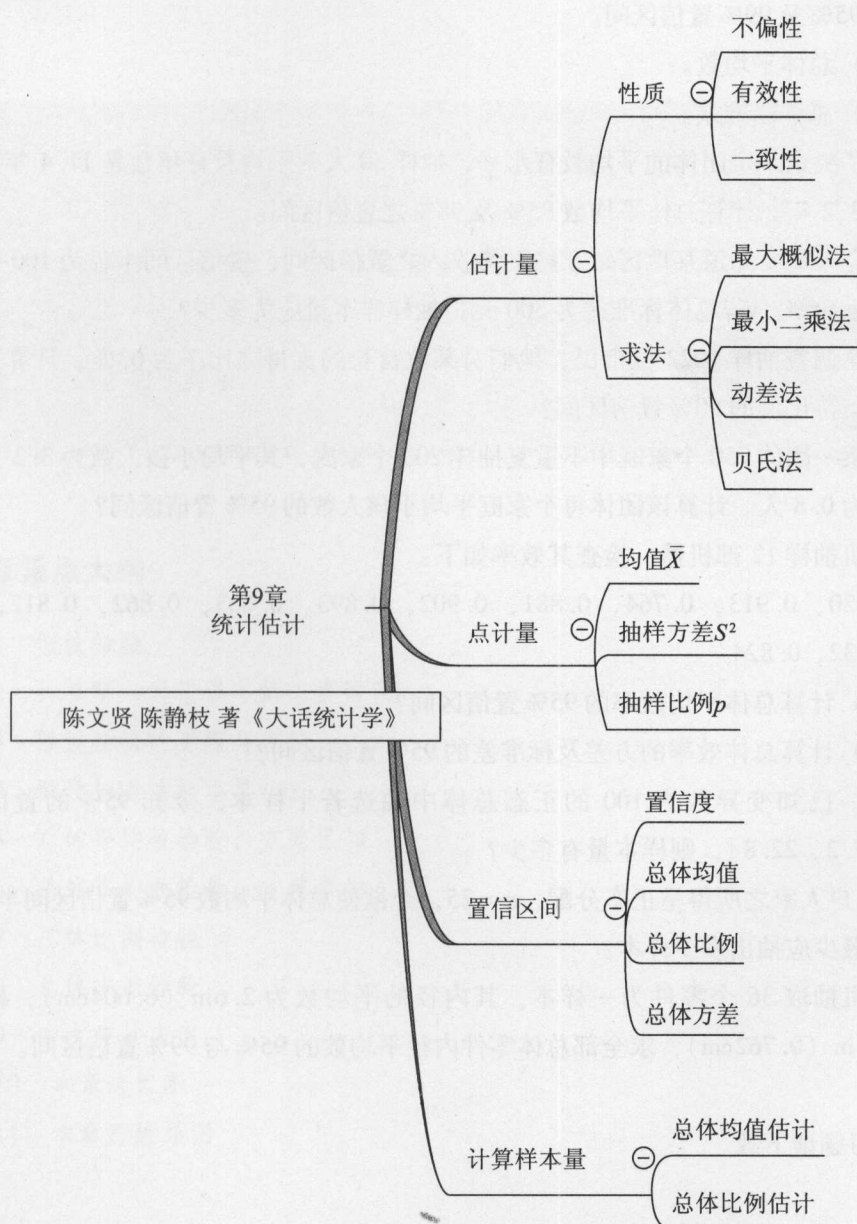


图 9-13 第9章思维导图

习题

1. 随机抽样 30 个 GRE 成绩, 平均分数为 1082 分, 标准差为 108 分, 决定下列参数的 95% 及 99% 置信区间。
 - (1) 总体平均数。
 - (2) 总体标准差。
2. 为了决定某个团体的平均教育水平, 抽样 50 人, 平均教育年数是 11.4 年, 标准差为 3.2 年。计算总体平均数 95% 及 98% 之置信区间。
3. 研究人员要知道某地区每家庭平均收入之置信区间, 置信区间半径为 100 元, 置信度为 95%, 若总体标准差为 500 元, 抽样样本量应为多少?
4. 民意调查抽样 1022 位市民, 他们对某位首长的支持之比率为 63%, 计算这位首长受支持比例的 90% 置信区间?
5. 从某一团体 750 个家庭中不重复抽样 200 个家庭, 其平均小孩人数为 3.2 人, 标准差为 0.8 人。计算该团体每个家庭平均小孩人数的 95% 置信区间?
6. 随机抽样 12 部机器, 检查其效率如下:
0.820, 0.913, 0.764, 0.881, 0.902, 0.893, 0.663, 0.862, 0.812, 0.778, 0.932, 0.824
 - (1) 计算总体平均效率的 95% 置信区间?
 - (2) 计算总体效率的方差及标准差的 95% 置信区间?
7. 从一已知变异数为 100 的正态总体中抽选若干样本, 今知 95% 的置信区间为 $[17.2, 22.8]$, 则样本量有多少?
8. 500 户人家之所得呈正态分配, $\sigma = 25$, 今欲使总体平均数 95% 置信区间半径为 10, 则最少应抽出多少样本?
9. 随机抽取 36 个零件为一样本, 其内径的平均数为 2.6in (6.604cm), 标准差为 0.3in (0.762cm), 求全部总体零件内径平均数的 95% 与 99% 置信区间。

其他习题请下载。



第10章

统计检验

无用之辩，不急之察，弃而不治。

——《荀子·天论篇》

二论各有所见，故是非曲直，未有所定。

——汉·王充《论衡》

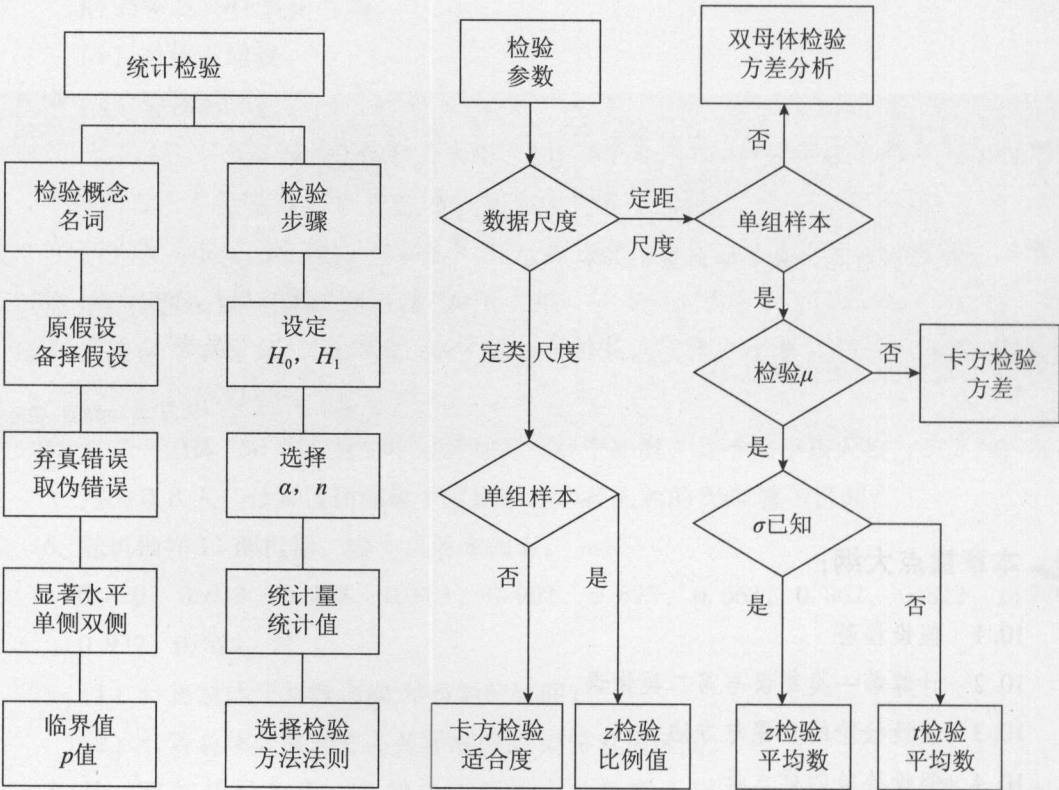
上之不明，因生辩也。

——韩非子



本章重点大纲：

- 10.1 假设检验
- 10.2 计算第一类错误与第二类错误
- 10.3 假设检验的步骤与方法
- 10.4 假设检验的样本量
- 10.5 总体平均数检验，方差已知
- 10.6 总体平均数检验，方差未知
- 10.7 总体比例检验
- 10.8 总体方差检验
- 10.9 中文统计应用
- 10.10 本章流程图
- 10.11 本章思维导图



本章概念图

10.1 假设检验

假定 (assumption): 使用统计方法或模型, 数据需要符合的条件, 例如: 总体是正态分布或两总体的方差相等。

假设 (hypothesis): 想要验证有关总体特征值 (参数) 的叙述, 例如: 总体不良率小于 2% 或两总体的平均数相等。

假设检验 (hypothesis testing) 是先对总体参数做一个叙述, 称作统计假设 (hypothesis), 然后利用抽样试验的样本数据, 来判断这个假设是否成立。

统计假设是一个对总体参数的假想答案, 实际上是“真”或“伪”, 我们事先不知道, 因为参数是未知的常数 (固定数)。所以, 我们只能用样本数据去决定是否“接受”或“拒绝”统计假设, 而这个决定可能是错误的, 这是统计假设的“误差”。

定义 假设检验的第一个假设, 称为原假设 (null hypothesis), 记作 H_0 。原假设的相反假设, 称为备择假设 (alternate hypothesis), 记作 H_1 。

例如对总体参数 θ , 平均数 μ , 或比例 π , 复合假设有 3 种型态:

$$\text{双侧检验: } \begin{cases} H_0: \theta = \theta_0 \\ H_1: \theta \neq \theta_0 \end{cases} \quad \begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases} \quad \begin{cases} H_0: \pi = \pi_0 \\ H_1: \pi \neq \pi_0 \end{cases}$$

$$\text{左侧检验: } \begin{cases} H_0: \theta \geq \theta_0 \\ H_1: \theta < \theta_0 \end{cases} \quad \begin{cases} H_0: \mu \geq \mu_0 \\ H_1: \mu < \mu_0 \end{cases} \quad \begin{cases} H_0: \pi \geq \pi_0 \\ H_1: \pi < \pi_0 \end{cases}$$

$$\text{右侧检验: } \begin{cases} H_0: \theta \leq \theta_0 \\ H_1: \theta > \theta_0 \end{cases} \quad \begin{cases} H_0: \mu \leq \mu_0 \\ H_1: \mu > \mu_0 \end{cases} \quad \begin{cases} H_0: \pi \leq \pi_0 \\ H_1: \pi > \pi_0 \end{cases}$$

原假设一定有等式 ($=, \leq, \geq$), 而且以后做检验依据的统计量, 是以原假设的等式, 推导检验法则, 有些统计学的教科书, 3 种检验 (双侧、左侧、右侧) 的原假设, 都定义为 $H_0: \theta = \theta_0$ 。

定义 假设检验会有一个决策法则 (decision rule), 当样本数据计算结果在某个范围, 则拒绝 (reject) 原假设; 否则, 接受 (accept), 较传统或严谨的说法是不拒绝 (do not reject) 原假设。

定义 假设检验有可能决策错误, 当原假设 H_0 实际上是真的, 但是根据样本数据计算结果, 却拒绝原假设, 这种错误称为第一类错误 (type I error) 或弃真错误。我们以

α 表示“第一类错误的概率”（简称第一类错误）：

$$\alpha = P \{ \text{拒绝 } H_0 \mid H_0 \text{ 为真} \} = \text{拒绝“对的 } H_0 \text{”的概率} = \text{“弃真错误”的概率}$$

定义 当备择假设 H_1 实际上是真的，但是根据样本数据计算结果，却接受原假设，这种错误称为第二类错误（type II error）或取伪错误。我们以 β 表示“第二类错误的概率”（简称第二类错误）：

$$\beta = P \{ \text{接受 } H_0 \mid H_1 \text{ 为真} \} = \text{接受“错的 } H_0 \text{”的概率} = \text{“取伪错误”的概率}$$

第一类错误 α 是检验的一个重要指标，称作假设检验的显著性水平（significance level）。

显著性水平是第一类错误的上限。

推断统计的统计检验，多数希望得到“显著”的结果，就是拒绝原假设。

(1) 总体参数检验：

拒绝原假设 \rightarrow 显著结果 \rightarrow 总体参数不等于假（预）设值 \rightarrow 不符规格 H_0

(2) 两个总体参数检验：

拒绝原假设 \rightarrow 显著结果 \rightarrow 两个总体参数有不相等

(3) 两个以上总体定距变量参数检验：

拒绝原假设 \rightarrow 显著结果 \rightarrow 各总体存在差异

(4) 两个以上定距变量检验：

拒绝原假设 \rightarrow 显著结果 \rightarrow 变量有（线性）相关

(5) 一个定类变量检验：

拒绝原假设 \rightarrow 显著结果 \rightarrow 总体不符合某个概率分布





(6) 两个定类变量检验：

拒绝原假设 \rightarrow 显著结果 \rightarrow 变量不是独立

以上 (2), (3), (4), (6) 的显著结果，表示总体的分类（因），有因果关系的显著影响。

如表 10-1，显著的结果，不是大好就大坏，所以要降低大坏（第一类错误）的概率 α ，或者说控制第一类错误的概率 α 在可接受的程度（通常是 0.05）以下。

表 10-1 假设检验的错误关系

检验结果	实际	
	H_0 为真	H_1 为真
不拒绝 H_0 接受 H_0	正确决策 但非幸运, 因为检验结果不显著, 可能要有新 H_0 概率 = $1 - \alpha$ = 置信度 = P (接受 H_0 H_0 为真) 	第二类错误 (取伪错误) 错误但非致命, 风险或误差 概率无法控制 概率 = P (第二类错误) = β = P (接受 H_0 H_1 为真) 
拒绝 H_0 接受 H_1 (显著)	第一类错误 (弃真错误) 错误且后果非常严重, 所以要控 制这个错误概率 概率上限 = α = 显著性水平 = P (拒绝 H_0 H_0 为真) 	正确决策 很幸运, 可以说服对方或公 布结果 (显著) 概率 = $1 - \beta$ = 检验力 = P (拒绝 H_0 H_1 为真) 

10.1.1 根据第一类错误决定原假设

如果已经选择要检验的参数, 那么建立原假设有 3 种型态: 双侧检验, 左侧检验, 右侧检验。双侧检验比较没有争议, 有问题的: 是左侧检验或右侧检验到底选那一个? 因为第一类错误 α 是可以控制的, 所以我们要根据第一类错误 α 来决定原假设的型态。考虑下述 4 种情况。

1) 想要拒绝的假设, 当作原假设; 或者想要调查或研究的假设, 当作备择假设。

先假设 H_0 成立, 再去拒绝它 \Rightarrow 让对方心服。

例如: 你对成功率 (比例) π 的看法, 是大于 0.6, 而我的看法是小于 0.6。我要拒绝你的看法, 所以我检验 $H_0: \pi \geq 0.6$ 。这样一来, 如果检验结果, 是拒绝 H_0 , 则我可以说检验误差只有 α 。我们对“拒绝原假设”的信心越强, 则 α 可以设定越小。

2) 拒绝后错误成本较高的假设, 当作原假设。避免第一类错误 α 造成更大损失 \Rightarrow 使自己心安。

如果“第一类错误的成本”越高, 则 α 可以设定越小。

例如: 检验产品某一规格 (如: 质量, 长度) 的平均数是大于 100 或小于 100。如果实际上平均数是小于 100, 而我们却宣布大于 100, 则要负担很大的赔偿和诉讼成本; 反之, 如果实际上平均数是大于 100, 而我们却宣布小于 100, 则损失很小。所以, 原假设是 $H_0: \mu \leq 100$, 即

$$\alpha = P \{ \text{宣布 } \mu > 100 \mid \text{实际上 } \mu \leq 100 \}$$

$$\beta = P \{ \text{宣布 } \mu \leq 100 \mid \text{实际上 } \mu > 100 \}$$

3) 为了表示因果关系是否显著, 因素 (两总体、方差分析、回归) 的影响结果是否

有差异，当然“等号”是原假设（双侧检验）。如果问题是检验总体参数“大于、小于”是否“显著”，则其相反的叙述为原假设。

例如：检验产品某一规格（如：质量，长度）的标准差大于 20 是否显著？原假设是： $H_0: \sigma^2 \leq 100$ 。

4) 有些教科书是按照样本数据的结果，来建立原假设的形态。例如：样本数据是 $\bar{x} < \mu_0$ ，则原假设是 $H_0: \mu \geq \mu_0$ ，因为如果原假设是 $H_0: \mu \leq \mu_0$ ，则根本不必计算检验法则，直接就可以得到结论：接受 H_0 （见 10.3 节检验步骤）。这种说法是以考试的观点来建立原假设。我们建议还是以要控制的第一类错误 α 来建立原假设。

10.1.2 第一类错误例子

1) 质量管理，总体（整批零件）不良率（ p ）小于等于 3%，才整批通过，检验 $H_0: p \leq 3\%$ 。

第一类错误 α ：实际上是整批不良率小于等于 3%，但是抽样检验后却不通过，称为生产者风险（producer's risk）。

第二类错误 β ：实际上是整批不良率大于 3%，但是抽样检验后却通过，称为消费者风险（consumer's risk）。

你觉得哪一个错误比较严重？

2) 美国大学教师联谊会，抗议教师升正教授平均年数 15 年太长，校方主管为了拒绝这种说法，抽样检验 $H_0: \mu \geq 15$ 。

第一类错误 α ：实际平均年数大于 15 年，却拒绝它。

第二类错误 β ：实际平均年数小于 15 年，却承认大于 15 年。

3) 法院判决被告，如果以不冤枉无辜为优先原则， H_0 ：被告无罪， H_1 ：被告有罪。

第一类错误 α ：实际上被告无罪，但是却判定有罪。这是“冤狱”。

第二类错误 β ：实际上被告有罪，但是却无罪释放。这是“纵放”。

10.2 计算第一类错误与第二类错误

假设检验平均数 μ ，下列检验 3 种型态： H_0^{I} ， H_0^{II} ， H_0^{III} 。

$$\begin{array}{lll} \text{双侧检验: } \begin{cases} H_0^{\text{I}}: \mu = \mu_0 \\ H_1^{\text{I}}: \mu \neq \mu_0 \end{cases} & \text{左侧检验: } \begin{cases} H_0^{\text{II}}: \mu \geq \mu_0 \\ H_1^{\text{II}}: \mu < \mu_0 \end{cases} & \text{右侧检验: } \begin{cases} H_0^{\text{III}}: \mu \leq \mu_0 \\ H_1^{\text{III}}: \mu > \mu_0 \end{cases} \end{array}$$

定义 拒绝域法（或称临界值法）的决策法则，是接受 H_0 的区域，称作接受域（acceptance region）。接受域以外的区域称作拒绝域（rejection region），是拒绝 H_0 。

定义 H_0^I ：拒绝 H_0 的区域 $(-\infty, x_L) \cup (x_U, +\infty)$ ，在 μ_0 的两侧，称为双侧检验（two-tailed test），如图 10-1（a）所示。

定义 H_0^{II} ：拒绝 H_0 的区域 $(-\infty, x_L)$ ，在 μ_0 的左侧，称为左侧检验（left-tailed test），如图 10-1（a）所示。

定义 H_0^{III} ：拒绝 H_0 的区域 $(x_U, +\infty)$ ，在 μ_0 的右侧，称为右侧检验（right-tailed test），如图 10-1（c）所示。拒绝域有的又称为弃却域。

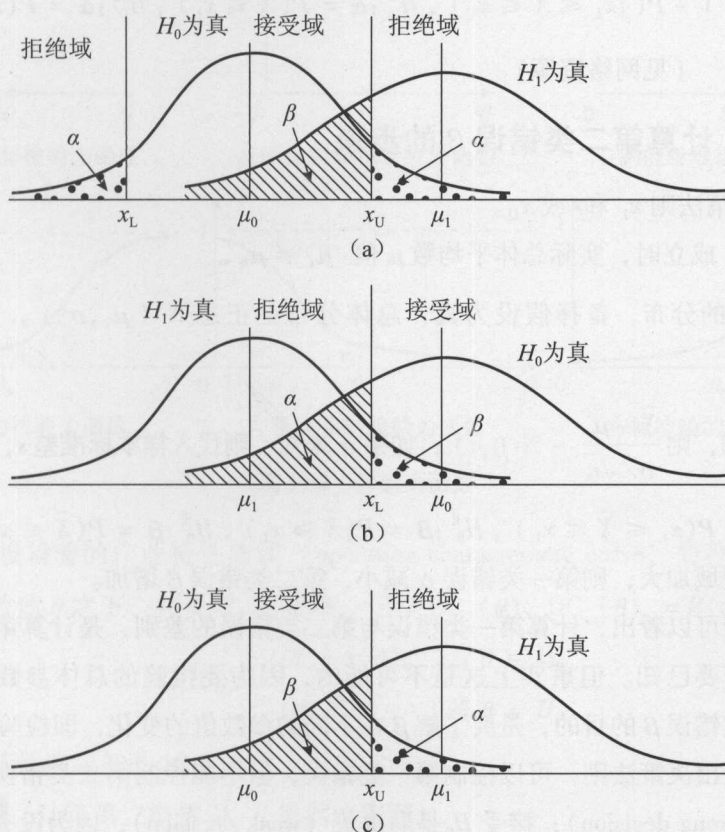


图 10-1 3 种检验型态的第一类错误与第二类错误

（a）双侧检验；（b）左侧检验；（c）右侧检验

上述之决策法则： x_L, x_U 称为临界值（critical values）。

接受域一定包括，原假设中的区域。例如：双侧检验的接受域 $\{\mu_0\} \subseteq (x_L, x_U)$ ， x_L ， x_U 在 μ_0 两边；左侧检验的接受域 $(\mu_0, \infty) \subseteq (x_L, \infty)$ ， x_L 在 μ_0 左边， $x_L \leq \mu_0$ ；右侧检验

的接受域 $(-\infty, \mu_0) \subseteq (-\infty, x_U)$, x_U 在 μ_0 右边, $\mu_0 \leq x_U$ 。

10.2.1 计算第一类错误 α 的步骤

- 1) 已知决策法则 x_L 和/或 x_U , 与要检验参数的统计量 (例如: μ 的统计量是 \bar{X})。
- 2) 决定 \bar{X} 的分布, 假设总体分布是正态 $N(\mu_0, \sigma^2)$ (这是原假设), 则 $\bar{X} \sim N(\mu_0, \frac{\sigma^2}{n})$ 。

如果 σ 已知, 则 $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$; 如果 σ 未知, 则代入样本标准差 S , $\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$ 。

- 3) $H_0^I: \alpha = 1 - P(x_L \leq \bar{X} \leq x_U)$, $H_0^{II}: \alpha = P(\bar{X} \leq x_L)$, $H_0^{III}: \alpha = P(\bar{X} \geq x_U)$ 。

例题 10.1 (见网络资源)

10.2.2 计算第二类错误 β 的步骤

- 1) 已知决策法则 x_L 和/或 x_U 。
- 2) 已知 H_1 成立时, 实际总体平均数 μ 值: $\mu_1 \neq \mu_0$ 。
- 3) 决定 \bar{X} 的分布, 备择假设为真, 总体分布是正态 $N(\mu_1, \sigma^2)$, 则 $\bar{X} \sim N(\mu_1, \sigma^2/n)$ 。

如果 σ 已知, 则 $\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} \sim N(0, 1)$; 如果 σ 未知, 则代入样本标准差 s , $\frac{\bar{X} - \mu_1}{s/\sqrt{n}} \sim t_{n-1}$ 。

- 4) $H_0^I: \beta = P(x_L \leq \bar{X} \leq x_U)$, $H_0^{II}: \beta = P(\bar{X} \geq x_L)$, $H_0^{III}: \beta = P(\bar{X} \leq x_U)$ 。

结论: 接受域加大, 则第一类错误 α 减小, 第二类错误 β 增加。

从以上步骤可以看出, 计算第一类错误与第二类错误的差别, 是计算第二类错误, 总体参数的实际值要已知。但事实上这是不可能的, 因为要检验的总体参数是未知的。但是, 计算第二类错误 β 的目的, 是要了解 β 在不同的参数值的变化, 即检验力和 OC 曲线。

所以我们定出决策法则, 可以控制第一类错误, 但不能控制第二类错误。因此: 拒绝 H_0 是强决策 (strong decision); 接受 H_0 是弱决策 (weak decision), 因为没有足够的证据来拒绝它。拒绝 H_0 , 尤其是双侧检验, 通常称为“显著”。

定义 假设检验在各种不同参数值 $\theta \in H_0$ 之下的第一类错误:

$$\alpha(\theta) = P(\text{拒绝 } H_0 \mid \theta \in H_0), \quad \alpha = \max_{\theta \in H_0} \alpha(\theta)$$

定义 假设检验在各种不同参数值 $\theta \in H_1$ 之下的第二类错误:

$$\beta(\theta) = P(\text{接受 } H_0 \mid \theta \in H_1)$$

定义 假设检验的检验力函数 (power function), 是在各种不同参数值 θ 之下, 拒绝 H_0 的概率, 记作 $PF(\theta)$ 。检验力 (power) = $PF(\theta) = P(\text{拒绝 } H_0 \mid \theta)$

$$PF(\theta) = \gamma(\theta) = \begin{cases} \alpha(\theta), & \text{若 } \theta \in H_0 \\ 1 - \beta(\theta), & \text{若 } \theta \in H_1 \end{cases}$$

因为 α 是固定的, $\alpha(\theta) < \alpha, \forall \theta \in H_0$, 所以检验力函数主要是看 $1 - \beta(\theta)$ 检验力函数曲线越陡, 则 $\alpha(\theta)$ 与 $\beta(\theta)$ 越小, 检验力越好, 如图 10-2 所示。

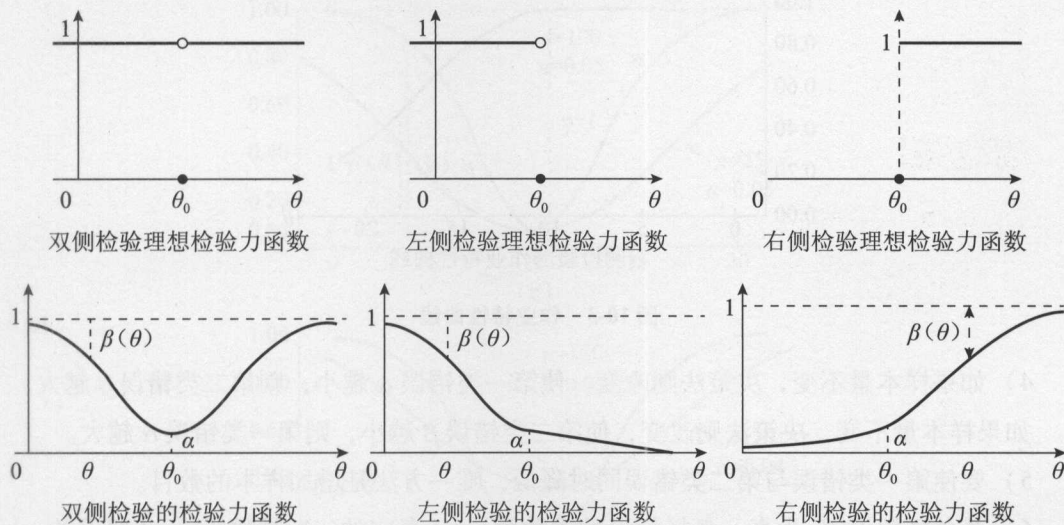


图 10-2 检验力函数

定义 假设检验的作业特性曲线 (operating characteristic curve, 简称 OC 曲线), 是在各种不同参数值 θ 之下, 接受 H_0 的概率, 记作 $OC(\theta)$ 。 $OC(\theta) = P(\text{接受 } H_0 \mid \theta)$

$$OC(\theta) = \begin{cases} 1 - \alpha(\theta), & \text{若 } \theta \in H_0 \\ \beta(\theta), & \text{若 } \theta \in H_1 \end{cases}$$

各类作业特性曲线如图 10-3 所示。

例题 10.2 计算第二类错误。(见网络资源)

以下是第一类错误与第二类错误的几个重要观念。

- 1) 根据决策法则与统计量 (如 \bar{X}) 的分布, 可以计算第一类错误 α 。
- 2) 计算第二类错误 β , 除了要已知决策法则与统计量的分布, 还要知道总体的备择假设实际参数值。
- 3) 如果接受 H_0 的区域越大, 则 α 越小。

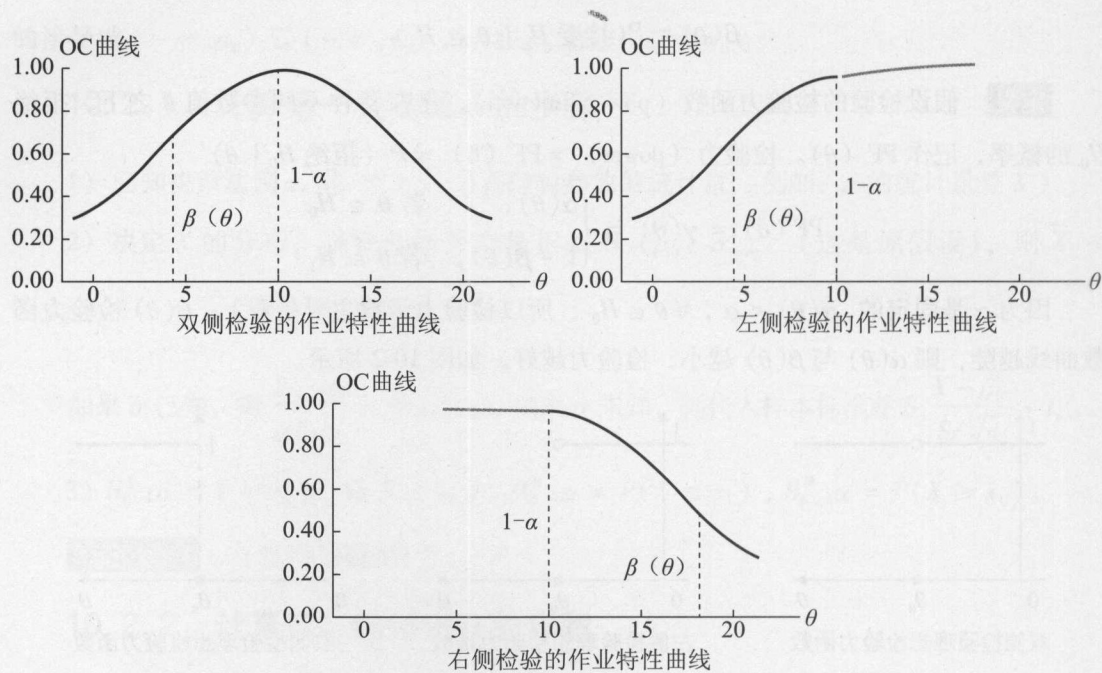


图 10-3 作业特性曲线

- 4) 如果样本量不变，决策法则改变，使第一类错误 α 越小，则第二类错误 β 越大。
如果样本量不变，决策法则改变，使第二类错误 β 越小，则第一类错误 α 越大。
- 5) 要使第一类错误与第二类错误同时降低，唯一方法是增加样本的数目。
- 6) 若备择假设 H_1 为真，备择假设的参数值 μ_1 (真实的 μ) 越接近 μ_0 (假设的 μ)，则第二类错误 β 越大。
- 7) 样本量、第一类错误、接受域长度是可以控制的变量 (请见 10.4 节)，第二类错误无法控制。

样本量、接受域、第一类错误、第二类错误四者之关系如图 10-4 所示。

图 10-4 (a) 双侧检验的检验力函数：样本量不变。若接受域减小，则第一类错误增加，第二类错误减小。若第一类错误增加，则接受域减小，第二类错误减小。

图 10-4 (b) 双侧检验的检验力函数：第一类错误不变。若样本量增加，则接受域减小，第二类错误减小。

图 10-4 (c) 双侧检验的检验力函数：接受域不变。若样本量增加，则第一类错误减小，第二类错误减小。(θ 接近 θ_0 例外)

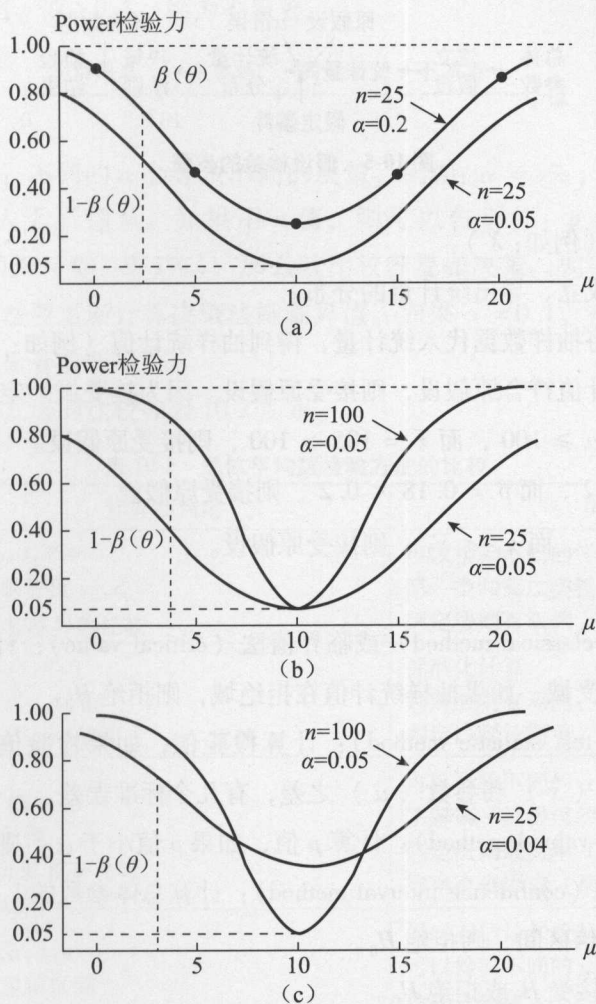


图 10-4 样本量、接受域、第一类错误、第二类错误四者之关系

10.3 假设检验的步骤与方法

1. 假设检验的步骤

假设检验的步骤如图 10-5 所示。

- 1) 了解问题，选出要检验的总体未知参数（例如： μ ）。
- 2) 建立原假设与备择假设。
- 3) 决定第一类错误（显著性水平） α ，选择样本量 n 。

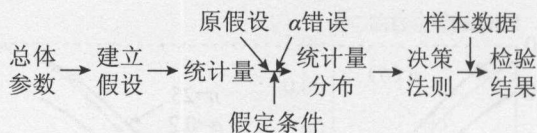


图 10-5 假设检验的步骤

4) 选择统计量 (例如: \bar{X})。

5) 利用原假设成立, 导出统计量的分布。

6) 进行抽样, 将抽样数据代入统计量, 得到抽样统计值 (例如: \bar{x})。

7) 如果抽样统计值符合原假设, 则接受原假设。因为接受域一定包括原假设。

例如: 检验 $H_0: \mu \geq 100$, 而 $\bar{x} = 105 > 100$, 则接受原假设。

检验 $H_0: \pi \leq 0.2$, 而 $p = 0.18 < 0.2$, 则接受原假设。

检验 $H_0: \mu_1 \leq \mu_2$, 而 $\bar{x}_1 < \bar{x}_2$, 则接受原假设。

2. 检验方法

(1) 拒绝域法 (classical method) 或临界值法 (critical value): 计算决策法则即临界值, 决定拒绝域及接受域, 如果抽样统计值在拒绝域, 则拒绝 H_0 。

(2) 检验值法 (test statistic method): 计算检验值, 如果检验值在拒绝域, 则拒绝 H_0 。检验值是统计量 (\bar{x}) 和参数 (μ) 之差, 有几个标准误差。

(3) p 值法 (p -valued method): 计算 p 值, 如果 p 值小于 α , 则拒绝 H_0 。

(4) 置信区间法 (confidence interval method): 计算总体参数的 $1 - \alpha$ 置信区间, 如果假定值 (μ_0) 不在置信区间, 则拒绝 H_0 。

得到检验结果: 接受 H_0 或拒绝 H_0 。

以上 4 种方法, 检验结果都相同, 拒绝域法与置信区间法的决策法则, 都是一个实数区间, 只是拒绝域法是以检验假设值 (如: μ_0) 为中心; 置信区间法是以抽样统计值 (如: \bar{x}) 为中心。

p 值 = $P \{ \text{拒绝 } H_0 \mid H_0 \text{ 为真, 且以抽样统计值为决策法则临界值} \}$

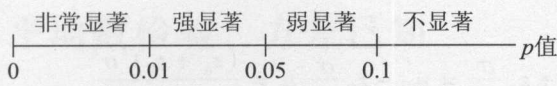
p 值是在原假设 H_0 为真的情况之下, 以抽样统计值为决策法则的第一类错误。 p 值是这个样本数据, 因为抽样误差, 造成拒绝 H_0 的错误判断的概率。

显著性水平 α 是第一类错误的上限。所以 p 值小于显著性水平 α , 表示以抽样统计值来拒绝原假设 H_0 的误差很小, 于是我们可以拒绝 H_0 。如果 p 值大于显著性水平 α , 表示以抽样统计值来拒绝原假设 H_0 的误差较大, 于是我们不能拒绝 H_0 。

p 值越低, 越可以拒绝 H_0 。

当 p 值 < 0.01 , 称为非常显著; 当 $0.01 < p$ 值 < 0.05 , 称为强显著; 当 $0.05 < p$ 值 $<$

0.1，称为弱显著；当 p 值 >0.1 ，称为不显著。



拒绝域法的缺点：不同的 α 会导致不同的决策，例如： $\alpha = 5\%$ ，接受 H_0 ； $\alpha = 10\%$ ，拒绝 H_0 。这样会使人无所适从。如果用 p 值，则可以告诉你： p 值是接近 5% （如： 5.5% ），或是接近 10% （如： 9.7% ），那么就比较容易做决策。换言之，如果显著性水平 α 改变，则拒绝域法要重新计算决策法则临界值。如果 $\alpha = 0.1$ ，拒绝域法不知道检验结果是非常显著、强显著或弱显著。

总体平均数检验方法的比较如表 10-2 所示。

表 10-2 总体平均数检验方法的比较

检验方法	计算和判定	优缺点
拒绝域法 (临界值法)	用 μ_0, α, σ, n 计算临界值 x_L, x_U → \bar{x} 和临界值比较	可以检验不同的样本组 \bar{x} 第一类和第二类错误的概念 判定法则有双侧、左右侧检验
检验值法	用 $\mu_0, \bar{x}, \sigma, n$ 计算 z^*, t^* 值 → z^*, t^* 值和 $z_{\alpha/2}, t_{\alpha/2}$ 比较	标准化计算 判定法则有双侧、左右侧检验 判定法则较简单
p 值法	用 z^* 值 计算 p 值 → p 值和 α 比较	可以检验不同的 α 了解第一类错误的程度 判定法则最简单（不分单双尾） 概率查表困难（最好用计算机）
置信区间法	用 $\bar{x}, \alpha, \sigma, n$ 计算置信区间 → μ_0 和置信区间比较	同时有估计和检验 可以检验不同的 μ_0 检验比例值 p 可能矛盾 卡方、无母数检验不适用

10.4 假设检验的样本量

要计算假设检验的样本量，必须先知道：第一类错误 α ，第二类错误 β ，总体标准差，以及总体的备择假设实际参数值。

10.4.1 总体均值检验的样本量

以右侧检验为例：

$$\begin{cases} H_0^{\text{III}}: \mu \leq \mu_0 \\ H_1^{\text{III}}: \mu > \mu_0 \end{cases}$$

假设总体是正态分布, 标准差 σ 已知, 第一类错误 α , 第二类错误 β , 总体的备择假设实际参数值 μ_1 。

$$\text{临界值 } x_U = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} = \mu_1 - z_\beta \frac{\sigma}{\sqrt{n}}, n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_1)^2}$$

$$\text{双侧检验的样本量: } n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_1)^2}$$

$$\text{左侧与右侧检验的样本量: } n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_0 - \mu_1)^2}$$

$$\text{所以 } z_\beta = \sqrt{\frac{n(\mu_0 - \mu_1)^2}{\sigma^2}} - z_\alpha$$

$$\text{双侧检验接受域的长度 } L = 2z_\alpha \frac{\sigma}{\sqrt{n}}$$

$\alpha, \beta, z_\alpha, z_\beta$ 的关系如下式所示。

$$\alpha \downarrow \Leftrightarrow z_\alpha \uparrow, \alpha \uparrow \Leftrightarrow z_\alpha \downarrow, \beta \downarrow \Leftrightarrow z_\beta \uparrow$$

从上述式子可以导出图 10-6 所示的关系 (“+” 表示两者有正相关, 同时增加或同时减少; “-” 表示两者有负相关, 一个增加则另一个减少)

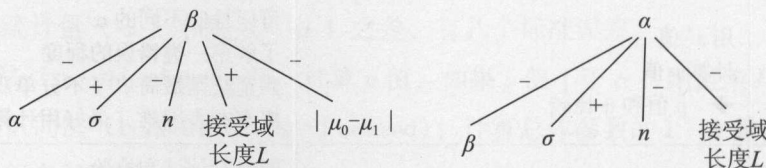


图 10-6 均值检验的样本量、接受域、第一类错误、第二类错误、总体标准差之关系

10.4.2 总体比例检验的样本量

$$\text{总体比例检验: } \begin{cases} H_0: p = p_0 \\ H_1: p \neq p_0 \end{cases}$$

假设第一类错误 α , 第二类错误 β , 总体的备择假设实际比例 p_1 。

$$\text{双侧检验的样本量: } n = \frac{[z_{\alpha/2} \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p_1(1-p_1)}]^2}{(p_0 - p_1)^2}$$

$$\text{左侧检验与右侧检验的样本量: } n = \frac{[z_\alpha \sqrt{p_0(1-p_0)} + z_\beta \sqrt{p_1(1-p_1)}]^2}{(p_0 - p_1)^2}$$

10.5 总体平均数检验，方差已知

假定条件：

- 1) 总体随机变量 X 为正态分布。 H_0 为真, $X \sim N(\mu_0, \sigma^2)$ 。(若总体不是正态, 则样本量 $n > 30$)
 - 2) 总体的标准差 σ 已知。 $\bar{X} \sim N(\mu_0, \frac{\sigma^2}{n})$ 。
 - 3) n 为样本的数目, \bar{x} 为样本平均数。
- 决策法则如表 10-3 所示。

表 10-3 决策法则（总体均值检验，方差已知）

决策 法则	双侧检验 $\begin{cases} H_0:\mu = \mu_0 \\ H_1:\mu \neq \mu_0 \end{cases}$	左侧检验 $\begin{cases} H_0:\mu \geq \mu_0 \\ H_1:\mu < \mu_0 \end{cases}$	右侧检验 $\begin{cases} H_0:\mu \leq \mu_0 \\ H_1:\mu > \mu_0 \end{cases}$
拒绝 域法	$x_L = \mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ $x_U = \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ 若 $\bar{x} \leq x_L$ 或 $\bar{x} \geq x_U$, 则拒绝 H_0	$x_L = \mu_0 - z_{\alpha} \frac{\sigma}{\sqrt{n}}$ 若 $\bar{x} \leq x_L$, 则拒绝 H_0	$x_U = \mu_0 + z_{\alpha} \frac{\sigma}{\sqrt{n}}$ 若 $\bar{x} \geq x_U$, 则拒绝 H_0
检验 值法	$z^* = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$ 若 $ z^* \geq z_{\alpha/2}$, 则拒绝 H_0	$z^* = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$ 若 $z^* \leq -z_{\alpha}$, 则拒绝 H_0	$z^* = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$ 若 $z^* \geq z_{\alpha}$, 则拒绝 H_0
p 值 法	p 值 = $2P(Z \geq z^*)$ 若 p 值 < α , 则拒绝 H_0	p 值 = $P(Z \leq z^*)$ 若 p 值 < α , 则拒绝 H_0	p 值 = $P(Z \geq z^*)$ 若 p 值 < α , 则拒绝 H_0
置信 区间 法	$\bar{x}_L = \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ $\bar{x}_U = \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ 若 $\mu_0 \leq \bar{x}_L$ 或 $\mu_0 \geq \bar{x}_U$, 则拒绝 H_0	$\bar{x}_U = \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}}$ 若 $\mu_0 \geq \bar{x}_U$, 则拒绝 H_0	$\bar{x}_L = \bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}}$ 若 $\mu_0 \leq \bar{x}_L$, 则拒绝 H_0

例题 10.3 （见网络资源）

例题 10.4 （见网络资源）

10.6 总体平均数检验，方差未知

假定条件：

- 1) 总体为正态分布。 H_0 为真, $X \sim N(\mu_0, \sigma^2)$ (若总体不是正态, 则样本量 $n > 30$)。
- 2) 总体的标准差 σ 未知。 $\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n - 1)$ 。
- 3) n 为样本的数目, \bar{x} 为样本平均数, s 为样本标准差。

决策法则如表 10-4 所示。

表 10-4 决策法则（总体均值检验，方差未知）

决策 法则	双侧检验 $\begin{cases} H_0:\mu = \mu_0 \\ H_1:\mu \neq \mu_0 \end{cases}$	左侧检验 $\begin{cases} H_0:\mu \geq \mu_0 \\ H_1:\mu < \mu_0 \end{cases}$	右侧检验 $\begin{cases} H_0:\mu \leq \mu_0 \\ H_1:\mu > \mu_0 \end{cases}$
拒绝 域法	$x_L = \mu_0 - t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}}$ $x_U = \mu_0 + t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}}$ 若 $\bar{x} \leq x_L$ 或 $\bar{x} \geq x_U$, 则拒 绝 H_0	$x_L = \mu_0 - t_{\alpha}(n - 1) \frac{s}{\sqrt{n}}$ 若 $\bar{x} \leq x_L$, 则拒绝 H_0	$x_U = \mu_0 + t_{\alpha}(n - 1) \frac{s}{\sqrt{n}}$ 若 $\bar{x} \geq x_U$, 则拒绝 H_0
检验 值法	$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ 若 $ t^* \geq t_{\alpha/2}(n - 1)$, 则拒绝 H_0	$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ 若 $t^* \leq -t_{\alpha}(n - 1)$, 则拒 绝 H_0	$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ 若 $t^* \geq t_{\alpha}(n - 1)$, 则拒绝 H_0
p 值 法	p 值 = $2P(t(n - 1) \geq t^*)$ 若 p 值 < α , 则拒绝 H_0	p 值 = $P[t(n - 1) \leq t^*]$ 若 p 值 < α , 则拒绝 H_0	p 值 = $P[t(n - 1) \geq t^*]$ 若 p 值 < α , 则拒绝 H_0
置信 区间 法	$\bar{x}_L = \bar{x} - t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}}$ $\bar{x}_U = \bar{x} + t_{\alpha/2}(n - 1) \frac{s}{\sqrt{n}}$ 若 $\mu_0 \leq \bar{x}_L$ 或 $\mu_0 \geq \bar{x}_U$, 则拒绝 H_0	$\bar{x}_U = \bar{x} + t_{\alpha}(n - 1) \frac{s}{\sqrt{n}}$ 若 $\mu_0 \geq \bar{x}_U$, 则拒绝 H_0	$\bar{x}_L = \bar{x} - t_{\alpha}(n - 1) \frac{s}{\sqrt{n}}$ 若 $\mu_0 \leq \bar{x}_L$, 则拒绝 H_0

例题 10.5 (见网络资源)

10.7 总体比例检验

假定条件:

- 1) 总体 X 为贝努里分布 $X \sim \text{Bern}(\pi_0)$, π 是成功率, π_0 是原假设的成功率。
- 2) n 是抽样的数目, T 是抽样的成功次数, $p = T/n$ 是样本的成功率。
- 3) 大样本, $n\pi_0 > 5$ 且 $n(1 - \pi_0) > 5$ 。
- 4) 统计量为 $p = \frac{T}{n}$, 若 n 相当大, 则 p 近似正态分布。 $p \sim N\left[\pi_0, \frac{\pi_0(1 - \pi_0)}{n}\right]$ 。
- 5) t 是实际抽样的成功次数, $p = t/n$ 是实际抽样的成功率 (统计值)。

在表 10-7 所示的检验法中, 计算标准差因为拒绝域法利用 π_0 , 计算上下限; 置信区间法利用 p , 计算上下限。两者半径不同, 有可能造成不一致的结果。例如拒绝域法结论是接受 H_0 , 置信区间法结论是拒绝 H_0 。因为 p 落在接受域, 但是 π_0 不在置信区间。

表 10-5 决策法则 (总体比例检验)

决策法则	双侧检验 $\begin{cases} H_0: \pi = \pi_0 \\ H_1: \pi \neq \pi_0 \end{cases}$	左侧检验 $\begin{cases} H_0: \pi \geq \pi_0 \\ H_1: \pi < \pi_0 \end{cases}$	右侧检验 $\begin{cases} H_0: \pi \leq \pi_0 \\ H_1: \pi > \pi_0 \end{cases}$
拒绝域法	$p_L = \pi_0 - z_{\alpha/2} \sqrt{\frac{\pi_0(1 - \pi_0)}{n}}$ $p_U = \pi_0 + z_{\alpha/2} \sqrt{\frac{\pi_0(1 - \pi_0)}{n}}$ 若 $p \leq p_L$ 或 $p \geq p_U$, 则拒绝 H_0	$p_L = \pi_0 - z_{\alpha} \sqrt{\frac{\pi_0(1 - \pi_0)}{n}}$ 若 $p \leq p_L$, 则拒绝 H_0	$p_U = \pi_0 + z_{\alpha} \sqrt{\frac{\pi_0(1 - \pi_0)}{n}}$ 若 $p \geq p_U$, 则拒绝 H_0
检验值法	$z^* = \frac{p - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}$ 若 $ z^* \geq z_{\alpha/2}$, 则拒绝 H_0	$z^* = \frac{p - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}$ 若 $z^* \leq -z_{\alpha}$, 则拒绝 H_0	$z^* = \frac{p - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}$ 若 $z^* \geq z_{\alpha}$, 则拒绝 H_0
p 值法	$p \text{ 值} = 2P(Z \geq z^*)$ 若 $p \text{ 值} < \alpha$, 则拒绝 H_0	$p \text{ 值} = P(Z \leq z^*)$ 若 $p \text{ 值} < \alpha$, 则拒绝 H_0	$p \text{ 值} = P(Z \geq z^*)$ 若 $p \text{ 值} < \alpha$, 则拒绝 H_0
置信区间法	$\bar{p}_L = p - z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}}$ $\bar{p}_U = p + z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}}$ 若 $\pi_0 \leq \bar{p}_L$ 或 $\pi_0 \geq \bar{p}_U$, 则拒绝 H_0	$\bar{p}_U = p + z_{\alpha} \sqrt{\frac{p(1 - p)}{n}}$ 若 $\pi_0 \geq \bar{p}_U$, 则拒绝 H_0	$\bar{p}_L = p - z_{\alpha} \sqrt{\frac{p(1 - p)}{n}}$ 若 $\pi_0 \leq \bar{p}_L$, 则拒绝 H_0

请注意: p 和 p 值不要搞混, p 是样本比例, p 值是检验的概率。

例题 10.6 (见网络资源)

10.8 总体方差检验

- 假定条件：
- 1) 总体为正态分布。
 - 2) 总体的标准差未知。
 - 3) n 为样本的数目， s 为样本标准差。
- 决策法则如表 10-6 所示。

表 10-6 决策法则（总体方差检验）

决策法则	双侧检验 $\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 \neq \sigma_0^2 \end{cases}$	左侧检验 $\begin{cases} H_0: \sigma^2 \geq \sigma_0^2 \\ H_1: \sigma^2 < \sigma_0^2 \end{cases}$	右侧检验 $\begin{cases} H_0: \sigma^2 \leq \sigma_0^2 \\ H_1: \sigma^2 > \sigma_0^2 \end{cases}$
拒绝域法	$s_L^2 = \frac{\sigma_0^2 \chi_{1-\frac{\alpha}{2}, n-1}^2}{n-1}, s_U^2 = \frac{\sigma_0^2 \chi_{\frac{\alpha}{2}, n-1}^2}{n-1}$ 若 $s^2 \leq s_L^2$ 或 $s^2 \geq s_U^2$ ，则拒绝 H_0	$s_L^2 = \frac{\sigma_0^2 \chi_{1-\alpha, n-1}^2}{n-1}$ 若 $s^2 \leq s_L^2$ ，则拒绝 H_0	$s_U^2 = \frac{\sigma_0^2 \chi_{\alpha, n-1}^2}{n-1}$ 若 $s^2 \geq s_U^2$ ，则拒绝 H_0
检验值法	$\chi_*^2 = \frac{(n-1)s^2}{\sigma_0^2}$ 若 $\chi_*^2 \leq \chi_{1-\alpha/2}(n-1)$ ，或 $\chi_*^2 \geq \chi_{\alpha/2}(n-1)$ ，则拒绝 H_0	$\chi_*^2 = \frac{(n-1)s^2}{\sigma_0^2}$ 若 $\chi_*^2 \leq \chi_{1-\alpha, n-1}$ ，则拒绝 H_0	$\chi_*^2 = \frac{(n-1)s^2}{\sigma_0^2}$ 若 $\chi_*^2 \geq \chi_{\alpha, n-1}$ ，则拒绝 H_0
p 值法	p 值 = $2\min\{P(\chi_{n-1}^2 \leq \chi_*^2), P(\chi_{n-1}^2 \geq \chi_*^2)\}$ 若 p 值 < α ，则拒绝 H_0	p 值 = $P(\chi_{n-1}^2 \leq \chi_*^2)$ 若 p 值 < α ，则拒绝 H_0	p 值 = $P(\chi_{n-1}^2 \geq \chi_*^2)$ 若 p 值 < α ，则拒绝 H_0
置信区间法	$\bar{s}_L^2 = \frac{s^2(n-1)}{\chi_{\frac{\alpha}{2}, n-1}^2}, \bar{s}_U^2 = \frac{s^2(n-1)}{\chi_{1-\frac{\alpha}{2}, n-1}^2}$ 若 $\sigma_0^2 \leq \bar{s}_L^2$ 或 $\sigma_0^2 \geq \bar{s}_U^2$ ，则拒绝 H_0	$\bar{s}_U^2 = \frac{s^2(n-1)}{\chi_{1-\alpha, n-1}^2}$ 若 $\sigma_0^2 \geq \bar{s}_U^2$ ，则拒绝 H_0	$\bar{s}_L^2 = \frac{s^2(n-1)}{\chi_{\alpha, n-1}^2}$ 若 $\sigma_0^2 \leq \bar{s}_L^2$ ，则拒绝 H_0

例题 10.7 （见网络资源）

10.9 中文统计应用

10.9.1 t 检验（例题 10.5）

执行“t 检验：总体均值检验”的操作示意图和结果如图 10-7 和图 10-8 所示。

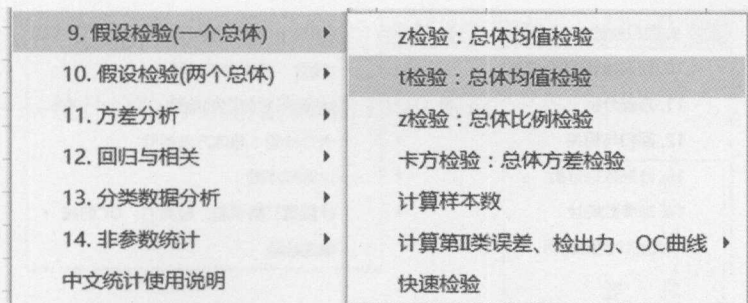


图 10-7 执行“t 检验：总体均值检验”的操作示意图

	A	B	C	D	E	F	G
1	t 检验：总体均值检验						
2							
3			<i>样本 1</i>				
4	显著性水平		0.05				
5	假设总体均值		846				
6	样本量		4				
7	样本均值		845				
8	样本标准差		1.414214		有限总体的数目		100
9	t 检验值		-1.41421		z 检验值		-1.43614
10							
11	单侧检验				单侧检验		
12	p 值		0.126108		p 值		0.123242
13	t 临界值		-2.35336		t 临界值		-2.35336
14							
15	双侧检验				双侧检验		
16	p 值		0.252215		p 值		0.246484
17	t 临界值		± 3.182446		t 临界值		± 3.182446

图 10-8 执行“t 检验：总体均值检验”的结果

10.9.2 计算第二类误差，检验力，OC 曲线

执行“计算第二类误差，检验力，OC 曲线”的操作示意图和结果如图 10-9 所示。

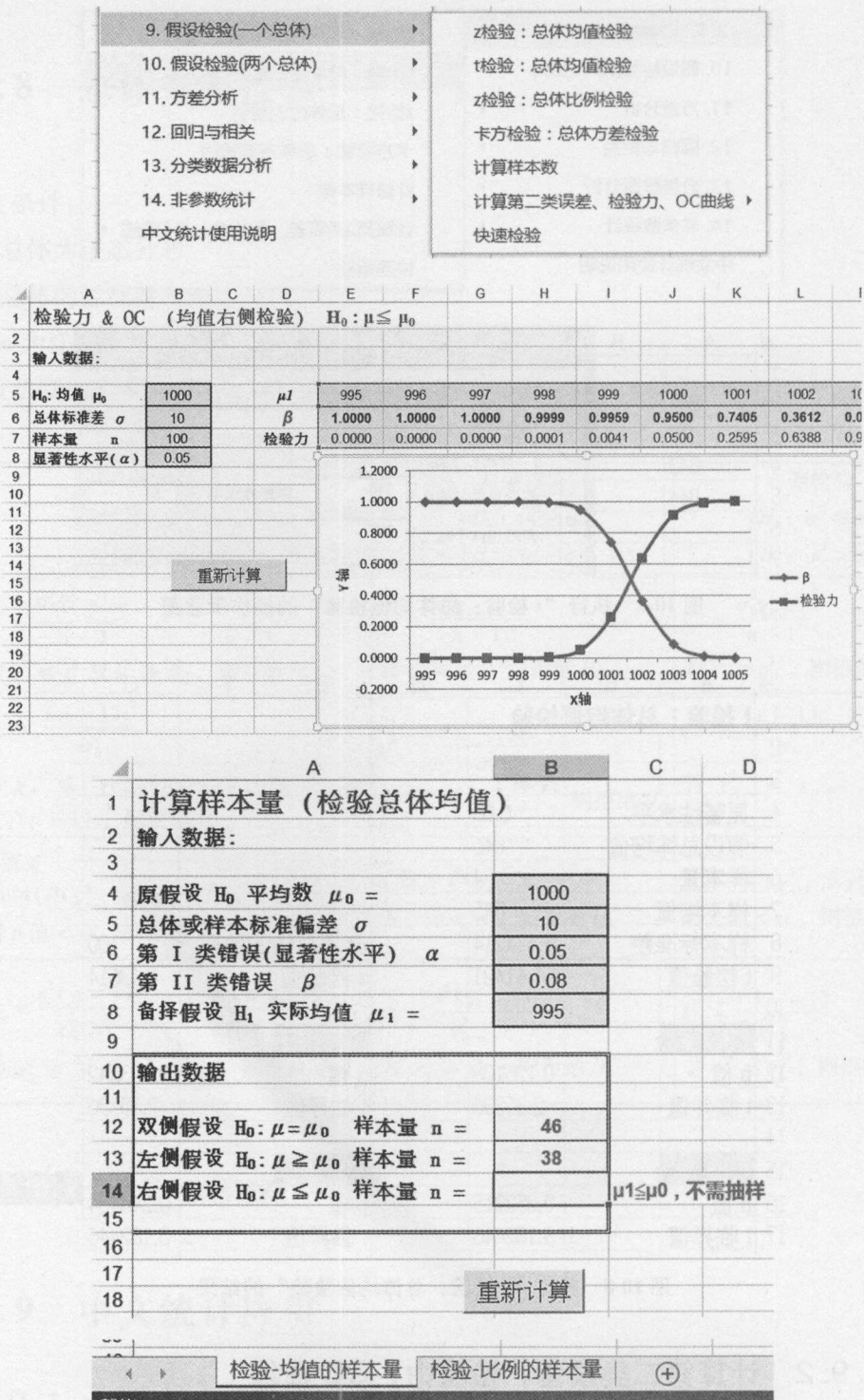


图 10-9 执行“计算第二类误差，检验力，OC 曲线”的操作示意图和结果

10.9.3 中文统计——统计检验的功能地图

中文统计——统计检验的功能地图如图 10-10 所示。

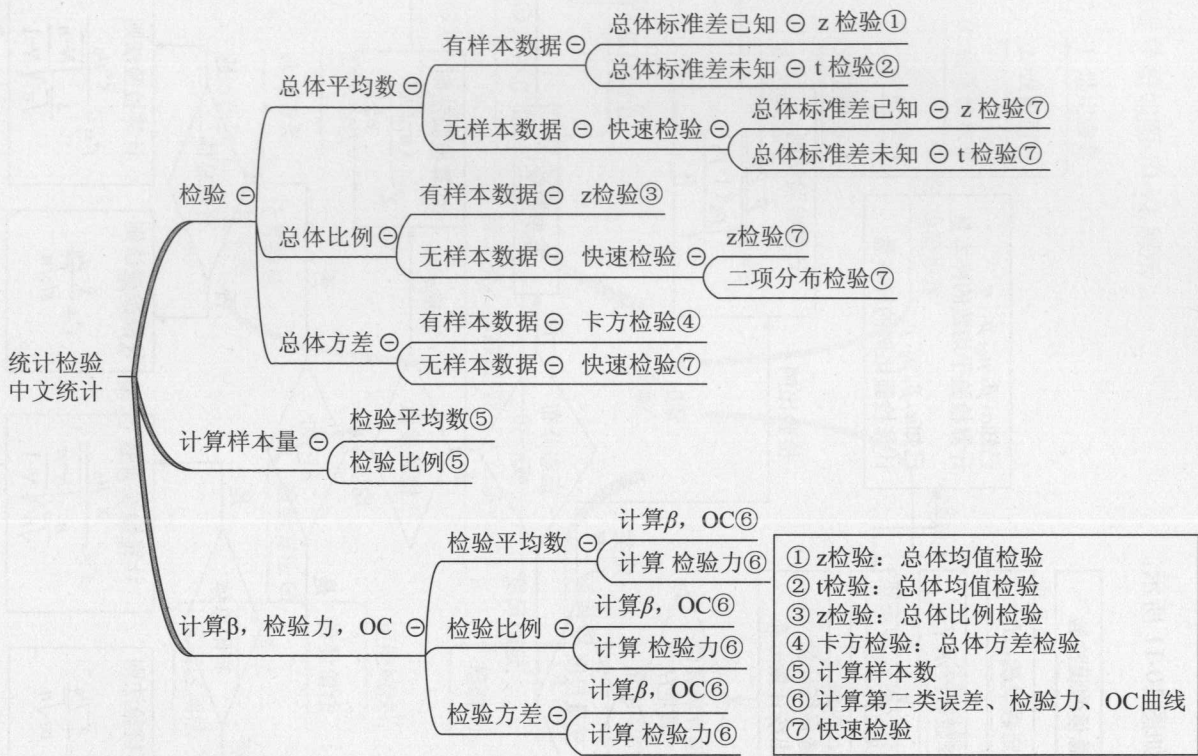


图 10-10 中文统计——统计检验的功能地图

10.10 本章流程图

本章流程图如图 10-11 所示。

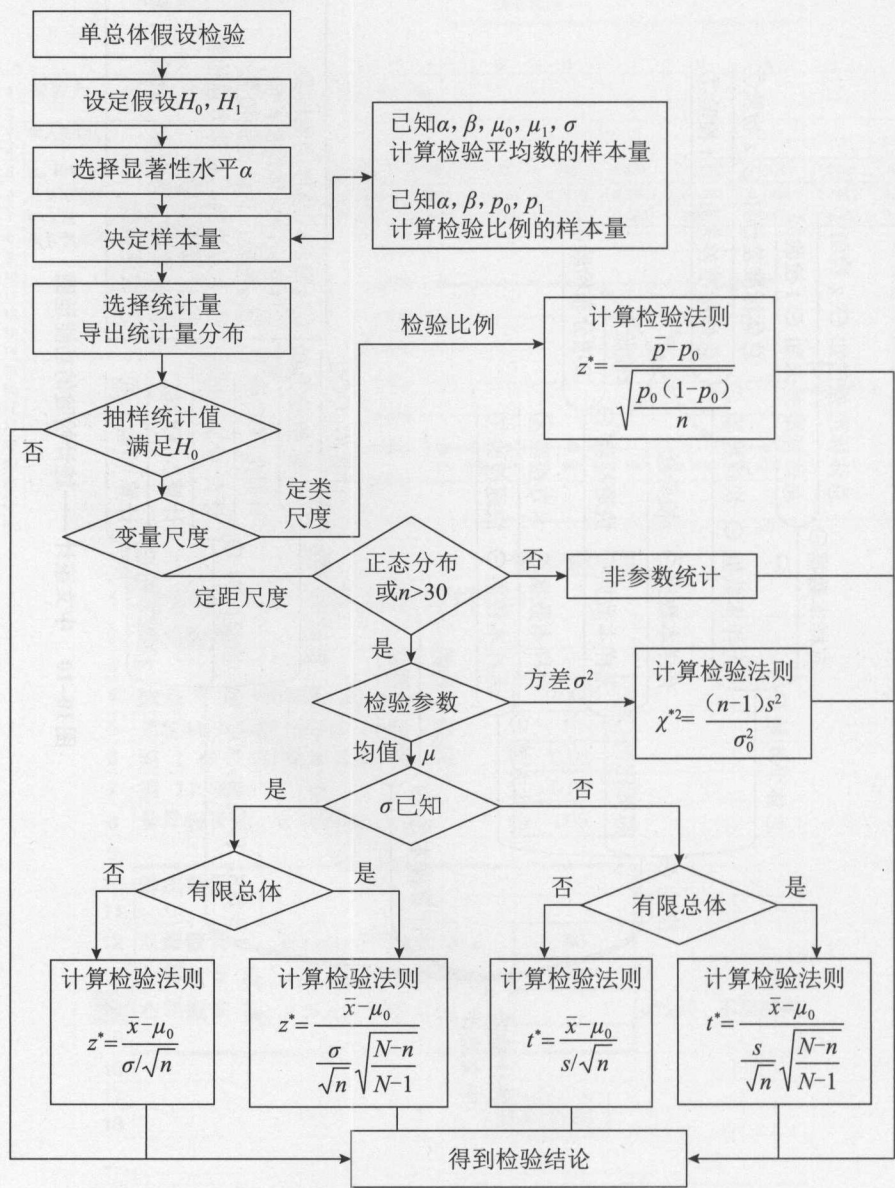


图 10-11 第 10 章流程图

10.11 本章思维导图

本章思维导图如图 10-12 所示。

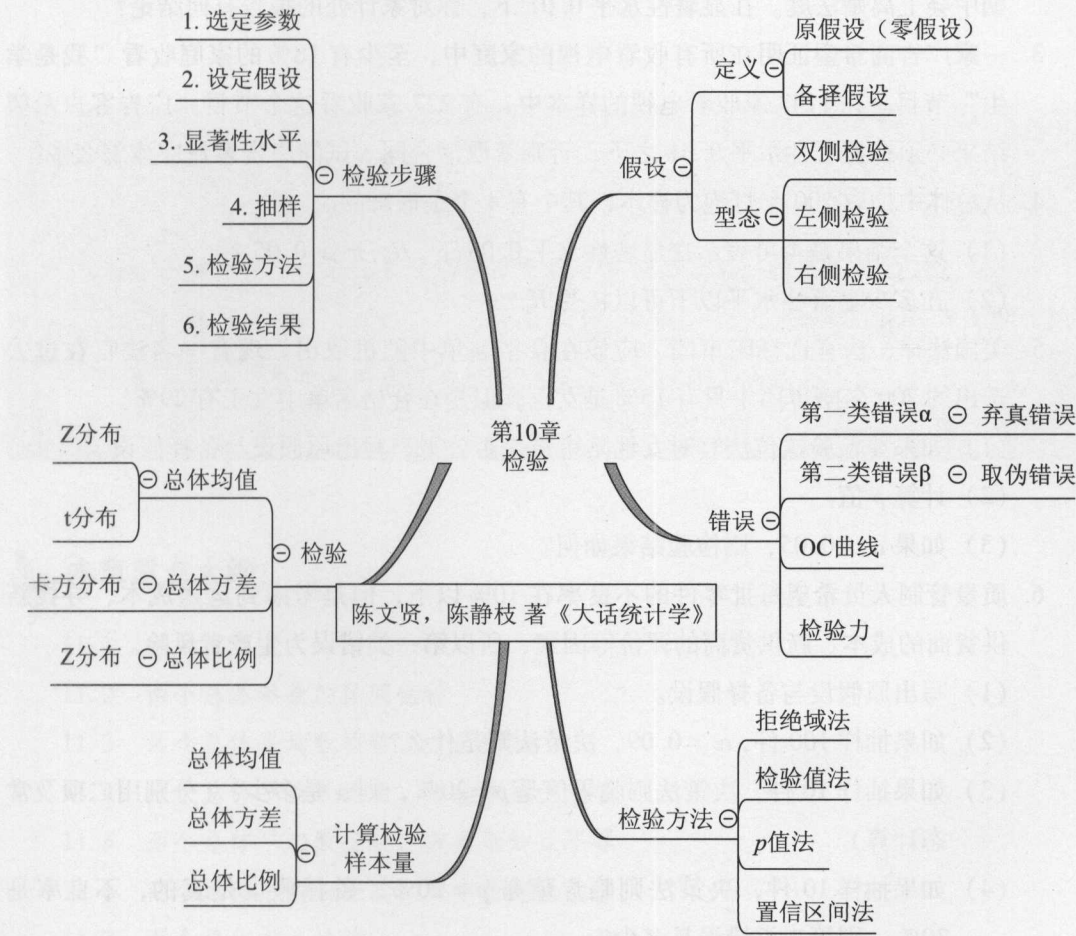


图 10-12 第 10 章思维导图

习 题

1. 在多少显著性水平下可以拒绝原假设 $H_0: \mu = 35$ 。
- 若 $\bar{x} = 37.4$, $s = 6.9$, $n = 60$, 其备择假设为:

- (1) $H_A: \mu > 35$ 。
 - (2) $H_A: \mu \neq 35$ 。
 - (3) $H_A: \mu < 35$ 。
2. 家计处报告指出在结婚一年后有 30% 以上的夫妇会上离婚法庭。一位怀疑这结论的统计学家发现, 在随机抽样 400 对结婚至少一年以上的夫妻, 108 对在第一年婚姻中会上离婚法庭。在显著性水平 0.01 下, 你对家计处的报告有何结论?
 3. 一家广告商希望证明在所有收看电视的家庭中, 至少有 18% 的家庭收看“我是学生”节目。在 1143 家收看电视的样本中, 有 227 家收看这个节目。广告客户希望结果必须在显著性水平 0.01 之下, 否则就取消合同。试问广告客户应该怎么做?
 4. 从总体中抽取 200 个灯泡为样本, 其中有 4 个是瑕疵品。
 - (1) 这个结果是否可表示在显著性水平 0.05 下, $H_0: \pi \geq 0.05$?
 - (2) 在多少显著性水平以下可以接受 H_0 ?
 5. 美国法律, 法官选择陪审团, 应该在合格名单中随机取出。现有一名法官在过去选出的 700 名陪审团中只有 15% 是女士, 但是在合格名单中女士有 29%。
 - (1) 如果要检验这位法官对女性陪审员是否公平, 写出原假设与备择假设。
 - (2) 计算 p 值。
 - (3) 如果 $\alpha = 0.05$, 则检验结果如何?
 6. 质量管制人员希望每批零件的不良率在 10% 以下, 但是考虑到运送成本、寻找新供货商的成本、新供货商的评价等因素, 所以第一类错误为生产者风险。
 - (1) 写出原假设与备择假设。
 - (2) 如果抽样 100 件, $\alpha = 0.09$, 决策法则是什么?
 - (3) 如果抽样 10 件, 决策法则临界值是 $p = 20\%$, 则 α 是多少? (分别用二项及常态计算)
 - (4) 如果抽样 10 件, 决策法则临界值是 $p = 20\%$, 备择假设是真的, 不良率是 30%, 则第二类错误是多少?
 - (5) 根据 (2) 的决策法则, 画出 OC 曲线。
 - (6) 根据 (3) 的决策法则, 画出 OC 曲线。

其他习题请下载。



第11章

两总体估计检验

欲穷千里目，更上一层楼。

——王之涣《登鹳雀楼》

人群分而物异产，来往贸迁以成宇宙。若各居而老死，何藉有群类哉？

——宋应星《天工开物》

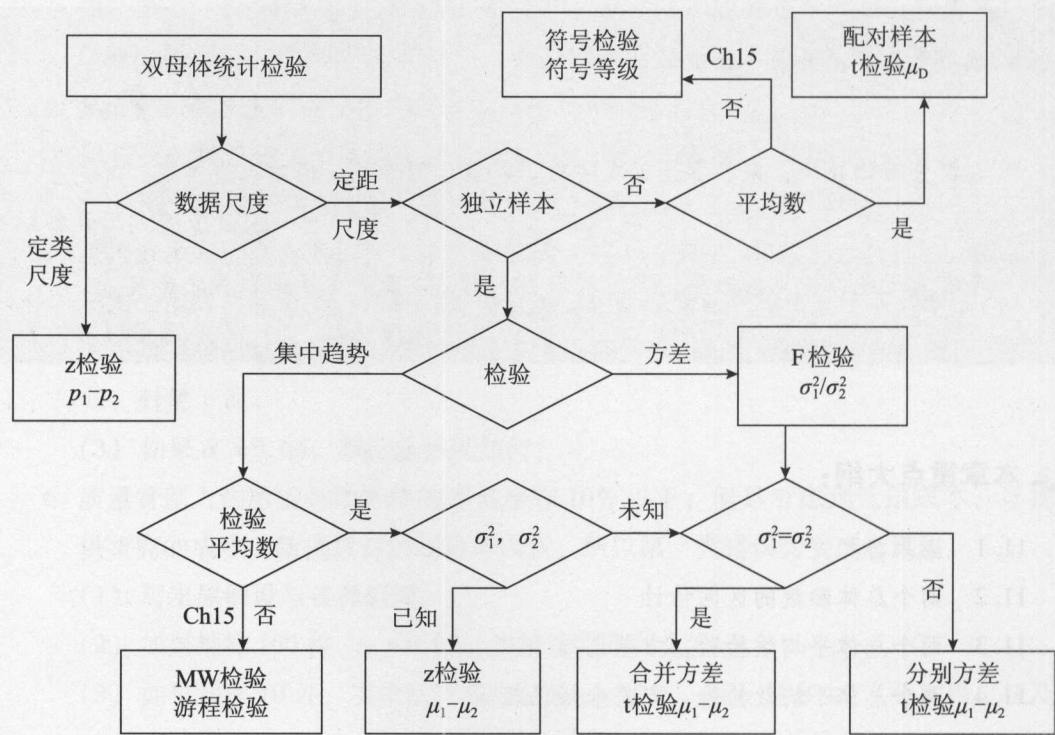
今有人于此，少见黑曰黑，多见黑曰白，则必以此人为不知白黑之辩矣。

——墨子《非攻》



本章重点大纲：

- 11.1 因果与相关
- 11.2 两个总体参数的区间估计
- 11.3 两个总体平均数检验，方差已知
- 11.4 两个总体平均数检验，方差未知但相等
- 11.5 两个总体平均数检验，方差未知且不等
- 11.6 两个总体平均数检验，样本是配对数据
- 11.7 两个总体方差检验
- 11.8 两个总体比例检验
- 11.9 中文统计应用
- 11.10 本章流程图
- 11.11 本章思维导图



本章概念图

11.1 因果与相关

我们在第一章提到因果关系，从本章开始，多数的统计方法，都是有因果或相关的分析。双总体统计检验，就是检验两个总体的参数（如：平均数，方差，比例）是否相等，或者相差一个常数。其因果关系，区分两个总体的变量，例如：性别、地区、时间、处理方法等，就是“因”；样本的观测或检验数据，就是“果”。其假设检验就是这个“因”对于“果”的影响是否显著（显著是备择假设成立，因素有影响，总体均值不相等）？所谓 A/B 检验（A/B test），例如：A 是实验组吃标靶药，B 是对照组（控制组）吃维他命，检验药物是否有效果（显著）。A/B 检验在网页设计，检验两个网页设计，对浏览者的点击率是否有差异。

要注意这两组样本是独立或配对。所谓两组样本是配对，即这两组样本在各总体是随机抽样，但是两组样本之间却有关。换句话说，第一组样本的某个抽样和第二组样本的某个抽样有关，而且两组样本量相等。例如：第一组样本是父亲的身高，第二组样本是儿子的身高。如果是独立样本，则是比较两代的平均身高；如果是配对样本，则第一组样本有一个父亲，第二组样本就一定有他的儿子，要比较的是父亲跟儿子之间的差异。

配对或“配对”样本的样本之间的“有关”，不是“变量”因果之间的“相关”，而是实验设计的“样本”关联。例如上述父子身高平均数的比较，上下代是“因”，身高是“果”，父子配对是“实验设计”。在方差分析的“区组”设计就是配对设计，扩大到两总体以上。

表 11-1 是表 1-3 增加独立与配对样本，这个表的流程图请见本章流程图。

表 11-1 统计检验的因果关系

统计方法		因果关系独立 或相关	一个变量（果）				多个 变量 因果	
			单母体 无因果 关系	另一个分类的独立变量（因）				
				独立 双母体	独立 多母体	配对 双样本		区组 多样本
有 参数 统计	检验 平均数	z 检验 t 检验	z 检验 t 检验	ANOVA	t 检验	随机区组 设计	回归 相关 分析	
	二项 比例值	z 检验	z 检验	χ^2 检验	McNemar 检验			
	检验 方差	χ^2 检验	F 检验	Hartley Bartlett			多变量 分析	

续表

统计方法 因果关系独立 或相关		一个变量（果）					多个 变量 因果
		单母体 无因果 关系	另一个分类的独立变量（因）				
			独立 双母体	独立 多母体	配对 双样本	区组 多样本	
非 参数 统计	检验 平均数	Sign 检验	χ^2 检验 Run 检验	KW 检验	Signed rank sum	Friedman 检验	Spearman 相关系数
	二项 比例值	二项 F 检验	超几何				
	多项 比例值	χ^2 检验	χ^2 检验	χ^2 检验			

11.2 两个总体参数的区间估计

双总体的参数区间估计，包括平均数差与比例差的区间估计。

11.2.1 两正态总体平均数差的区间估计

我们在统计的应用上，会比较两个总体的平均数，例如：①比较甲班与乙班的平均分；②比较实验组与观察组的平均生产数量；③比较制造业与服务业的劳工平均月薪。

假定两个总体均为正态分布，分别为 $N(\mu_1, \sigma_1^2)$ 与 $N(\mu_2, \sigma_2^2)$ ，或样本量大于等于 30。为了估计其平均数差 $\mu_1 - \mu_2$ ，分别从第一个总体抽样 n_1 个样本，得到样本平均数 \bar{x}_1 ，样本方差 s_1^2 ；从第二个总体抽样 n_2 个样本，得到样本平均数 \bar{x}_2 ，样本方差 s_2^2 。

现在分成几种情况，来说明平均数差 $\mu_1 - \mu_2$ 的置信区间估计。

1. 两组样本为独立抽样，总体方差 σ_1^2 及 σ_2^2 已知

在这种情况下，平均数差 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 置信区间为

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

如果置信区间的长度定为 L ，而第一个总体的抽样成本为每个样本 c_1 ，第二个总体的抽样成本为每个样本 c_2 。则两个总体的抽样数目 n_1, n_2 为

$$n_1 = 4 \left(\sigma_1^2 + \sigma_1 \sigma_2 \sqrt{\frac{c_2}{c_1}} \right) \left(\frac{z_{\alpha/2}}{L} \right)^2, \quad n_2 = 4 \left(\sigma_2^2 + \sigma_1 \sigma_2 \sqrt{\frac{c_1}{c_2}} \right) \left(\frac{z_{\alpha/2}}{L} \right)^2$$

如果总体标准差 σ_1 与 σ_2 未知，则上述样本量公式，可用预先抽样的样本标准差 s_1 与

s_2 代入。

如果是“不重复式抽样”（抽样后不放回），第一个总体的总体数目 N_1 ，抽样数目为 n_1 ，第二个总体的总体数目 N_2 ，抽样数目为 n_2 ，则

$$V(\bar{X}_1 - \bar{X}_2) = \frac{N_1 - n_1}{N_1 - 1} \times \frac{\sigma_1^2}{n_1} + \frac{N_2 - n_2}{N_2 - 1} \times \frac{\sigma_2^2}{n_2}$$

平均数差 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 置信区间的上下限为

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{V(\bar{X}_1 - \bar{X}_2)}$$

即

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{(N_1 - n_1)\sigma_1^2}{(N_1 - 1)n_1} + \frac{(N_2 - n_2)\sigma_2^2}{(N_2 - 1)n_2}}$$

2. 两组样本为独立抽样，总体方差 σ_1^2 及 σ_2^2 未知

在这种情况下，两总体平均数差的区间估计又分成两种情形计算。

1) 总体方差 σ_1^2 及 σ_2^2 未知，但是相等，则平均数差 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 置信区间为，将 \bar{x}_1, \bar{x}_2 代入下列式子：

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}(\nu) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}(\nu) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

合并方差

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

t 分布自由度

$$\nu = n_1 + n_2 - 2$$

2) 两组样本为独立抽样，总体方差 σ_1^2 及 σ_2^2 未知，但是不相等，则平均数差 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 置信区间为，将 \bar{x}_1, \bar{x}_2 代入下列式子：

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}(\nu) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}(\nu) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

t 分布自由度

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

3. 两组样本为配对样本

两组样本是配对，而且有配对的关系。第一组样本从第一个总体随机抽样，第二组样本和第一组样本有：时间上配对关系（如一前一后），血缘上配对关系（如父子兄弟）等

配对的关系。例如：

1) 同一个人的减肥前与减肥后的体重，减肥计划是因，体重是果，同一人是配对设计，消除个人因素。

2) 同一个人操作 A 机器与 B 机器的生产量，机器是因，产量是果，同一人是配对设计，消除个人因素。

3) 同一个家庭的父与子的身高，上下代是因，身高是果，同一家庭是配对设计，消除家庭因素。

4) 同一个体质（血型、年龄、过敏反应）的甲、乙两种药品的效果，药品是因，效果是果，同一体质是配对设计，消除体质因素。

5) 同一个班级（或智商）的两个考试的成绩；同班级两个人分别考两个考试，考试是因，成绩是果，同一班级是配对设计，消除班级（不同老师或学习气氛）因素。

置信区间的估计如下。

令第一个总体的样本数据为 x_{1i} ，第二个总体的样本数据为 x_{2i} 。两组样本数据的样本量一定相等，样本量为 n ，则计算

$$\begin{aligned}d_i &= x_{1i} - x_{2i} \\ \bar{x}_d &= \frac{\sum d_i}{n} = \bar{x}_1 - \bar{x}_2 \\ s_d^2 &= \frac{\sum (d_i - \bar{x}_d)^2}{n - 1}\end{aligned}$$

平均数差 $\mu_1 - \mu_2$ 的 $1 - \alpha$ 置信区间为，将 $\bar{x}_d, s_d = \sqrt{s_d^2}$ 代入下列式子，即

$$\bar{x}_d - t_{\alpha/2}(n-1) \frac{s_d}{\sqrt{n}} \leq \mu_1 - \mu_2 \leq \bar{x}_d + t_{\alpha/2}(n-1) \frac{s_d}{\sqrt{n}}$$

配对样本（例如：随机选出 5 个人比较上下两学期成绩）的置信区间，比起独立样本（例如：随机选出 5 个人的上学期成绩，另外随机选出 5 个人的下学期成绩）的置信区间，在相同的信赖度之下，应该是前者（配对样本）较小。

配对样本会使自由度减半（从 $2n - 2$ 变成 $n - 1$ ），损失自由度，浪费样本量，使 $t_{\alpha/2}$ 变大，增加置信区间的长度。但是因为配对样本可以消除许多外在的影响变量，例如：学生性别、智商、班级等外在变量除去后，配对样本只是单纯比较上下学期不同教材的学习程度。配对样本使方差减小，即 $s_d < s_p$ ， $s_p = \sqrt{V(X_1) + V(X_2)}$ 是独立样本 $X_1 - X_2$ 的标准差； $s_d = \sqrt{V(X_1 - X_2)}$ 是配对样本 $X_1 - X_2$ 的标准差。

由于

$$V(X_1 - X_2) = V(X_1) + V(X_2) - 2\text{Cov}(X_1, X_2), \text{Cov}(X_1, X_2) > 0$$

所以, 配对样本方差的减小抵得过自由度的损失。配对样本虽然自由度损失, 但可能消除许多外在的变量, 所以其置信区间的长度应该比独立样本的置信区间小。

11.2.2 两总体比例差的区间估计

我们在统计的应用上, 会比较两个总体的比例, 例如: ①比较甲班与乙班的及格比例; ②比较实验组与观察组的不良率; ③比较制造业与服务业的劳工投票率。

假定两个总体均为贝努里分布, 分别为 $\text{Bern}(\pi_1)$ 与 $\text{Bern}(\pi_2)$ 。为了估计其比例 (成功率) 差 $\pi_1 - \pi_2$, 分别从第一个总体抽样 n_1 个样本, 其中样本成功数目为 t_1 , 样本比例 (成功率) $p_1 = t_1/n_1$; 从第二个总体抽样 n_2 个样本, 其中样本成功数目为 t_2 , 样本比例 (成功率) $p_2 = t_2/n_2$ 。

假定两组样本为独立抽样, 比例差 $\pi_1 - \pi_2$ 的 $1 - \alpha$ 置信区间为

$$(p_1 - p_2) \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

如果是“不重复式抽样” (抽样后不放回), 第一个总体的总体数目 N_1 , 抽样数目为 n_1 , 第二个总体的总体数目 N_2 , 抽样数目为 n_2 , 则比例差 $p_1 - p_2$ 的 $1 - \alpha$ 置信区间为

$$(p_1 - p_2) \pm z_{\alpha/2} \sqrt{\frac{N_1 - n_1}{N_1 - 1} \times \frac{p_1(1-p_1)}{n_1} + \frac{N_2 - n_2}{N_2 - 1} \times \frac{p_2(1-p_2)}{n_2}}$$

有关上述置信区间的例题, 请见下列检验法则的置信区间法。

11.2.3 两总体方差比的区间估计

两个总体方差比的区间估计的假定条件:

- (1) 两组样本是独立, 而且分别从两个总体随机抽样。
- (2) 两个总体均为正态分布 (或样本量大于等于 30)。
- (3) 标准差未知 (当然未知, 才要估计)。
- (4) n_1, n_2 为样本量, s_1, s_2 为样本标准差。

$\frac{\sigma_1^2}{\sigma_2^2}$ 的 $1 - \alpha$ 置信区间:

$$\text{置信区间下限 } \bar{s}_L^2 = \frac{s_1^2}{s_2^2} F_{1-\alpha/2}(n_2 - 1, n_1 - 1)$$

$$\text{置信区间上限 } \bar{s}_U^2 = \frac{s_1^2}{s_2^2} F_{\alpha/2}(n_2 - 1, n_1 - 1)$$

11.3 两个总体平均数检验，方差已知

如果总体方差已知，双总体平均数检验的假定条件（如图 11-1 所示）：

- (1) 两组样本是独立的，而且分别从两个总体随机抽样。
- (2) 两个总体均为正态分布
(或样本量均大于等于 30, $n_1 \geq 30, n_2 \geq 30$)。
- (3) 两总体的标准差 σ_1, σ_2 都已知。
- (4) n_1, n_2 为样本量, \bar{x}_1, \bar{x}_2 为样本平均数。

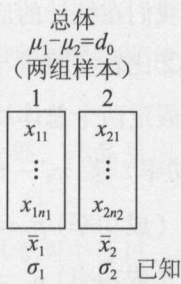


图 11-1 本节数据及假定条件

决策法则如表 11-2 所示。

表 11-2 决策法则（两个总体平均数检验，方差已知）

决策法则	双侧检验 $\begin{cases} H_0: \mu_1 - \mu_2 = d_0 \\ H_1: \mu_1 - \mu_2 \neq d_0 \end{cases}$	左侧检验 $\begin{cases} H_0: \mu_1 - \mu_2 \geq d_0 \\ H_1: \mu_1 - \mu_2 < d_0 \end{cases}$	右侧检验 $\begin{cases} H_0: \mu_1 - \mu_2 \leq d_0 \\ H_1: \mu_1 - \mu_2 > d_0 \end{cases}$
拒绝域法	$x_L = d_0 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ $x_U = d_0 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 若 $\bar{x}_1 - \bar{x}_2 \leq x_L$ 或 $\bar{x}_1 - \bar{x}_2 \geq x_U$ ， 则拒绝 H_0	$x_L = d_0 - z_{\alpha} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 若 $\bar{x}_1 - \bar{x}_2 \leq x_L$ ，则拒绝 H_0	$x_U = d_0 + z_{\alpha} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 若 $\bar{x}_1 - \bar{x}_2 \geq x_U$ ，则拒绝 H_0
检验值法	$z^* = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$ 若 $ z^* \geq z_{\alpha/2}$ ，则拒绝 H_0	$z^* = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$ 若 $z^* \leq -z_{\alpha}$ ，则拒绝 H_0	$z^* = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$ 若 $z^* \geq z_{\alpha}$ ，则拒绝 H_0
p 值法	p 值 = $2P(Z \geq z^*)$ 若 p 值 < α ，则拒绝 H_0	p 值 = $P(Z \leq z^*)$ 若 p 值 < α ，则拒绝 H_0	p 值 = $P(Z \geq z^*)$ 若 p 值 < α ，则拒绝 H_0
置信区间法	$\bar{x}_L = \bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ $\bar{x}_U = \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 若 $d_0 \leq \bar{x}_L$ 或 $d_0 \geq \bar{x}_U$ ，则拒绝 H_0	$\bar{x}_U = \bar{x}_1 - \bar{x}_2 + z_{\alpha} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 若 $d_0 \geq \bar{x}_U$ ，则拒绝 H_0	$\bar{x}_L = \bar{x}_1 - \bar{x}_2 - z_{\alpha} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ 若 $d_0 \leq \bar{x}_L$ ，则拒绝 H_0

例题 11.1 （见网络资源）

例题 11.2 （见网络资源）

11.4 两个总体平均数检验，方差未知但相等

总体方差未知但相等，双总体平均数检验的假定条件（如图 11-2 所示）：

- (1) 两组样本是独立，而且分别从两个总体随机抽样。
- (2) 两个总体均为正态分布（或样本量大于等于 30）。
- (3) 两总体的标准差未知但相等。
- (4) n_1, n_2 为样本量， \bar{x}_1, \bar{x}_2 为样本平均数， s_1, s_2 为样本标准差。

决策法则如表 11-3 所示。

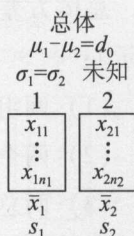


图 11-2 本节数据
及假定条件

表 11-3 决策法则（两个总体平均数检验，方差未知但相等）

决策法则	双侧检验 $\begin{cases} H_0: \mu_1 - \mu_2 = d_0 \\ H_1: \mu_1 - \mu_2 \neq d_0 \end{cases}$	左侧检验 $\begin{cases} H_0: \mu_1 - \mu_2 \geq d_0 \\ H_1: \mu_1 - \mu_2 < d_0 \end{cases}$	右侧检验 $\begin{cases} H_0: \mu_1 - \mu_2 \leq d_0 \\ H_1: \mu_1 - \mu_2 > d_0 \end{cases}$
拒绝域法	$x_L = d_0 - t_{\alpha/2}(\nu)s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $x_U = d_0 + t_{\alpha/2}(\nu)s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $\nu = n_1 + n_2 - 2$ $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$ 若 $\bar{x}_1 - \bar{x}_2 \leq x_L$ 或 $\bar{x}_1 - \bar{x}_2 \geq x_U$ ， 则拒绝 H_0	$x_L = d_0 - t_{\alpha}(\nu)s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $\nu = n_1 + n_2 - 2$ $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$ 若 $\bar{x}_1 - \bar{x}_2 \leq x_L$ ，则拒绝 H_0	$x_U = d_0 + t_{\alpha}(\nu)s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $\nu = n_1 + n_2 - 2$ $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$ 若 $\bar{x}_1 - \bar{x}_2 \geq x_U$ ，则拒绝 H_0
检验值法	$t^* = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{s_p \sqrt{(1/n_1) + (1/n_2)}}$ 若 $ t^* \geq t_{\alpha/2}(\nu)$ ，则拒绝 H_0	$t^* = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{s_p \sqrt{(1/n_1) + (1/n_2)}}$ 若 $t^* \leq -t_{\alpha}(\nu)$ ，则拒绝 H_0	$t^* = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{s_p \sqrt{(1/n_1) + (1/n_2)}}$ 若 $t^* \geq t_{\alpha}(\nu)$ ，则拒绝 H_0
p 值法	$p \text{ 值} = 2P[t(\nu) \geq t^*]$ 若 $p \text{ 值} < \alpha$ ，则拒绝 H_0	$p \text{ 值} = P[t(\nu) \leq t^*]$ 若 $p \text{ 值} < \alpha$ ，则拒绝 H_0	$p \text{ 值} = P[t(\nu) \geq t^*]$ 若 $p \text{ 值} < \alpha$ ，则拒绝 H_0
置信区间法	$\bar{x}_d = \bar{x}_1 - \bar{x}_2$ $\bar{x}_L = \bar{x}_d - t_{\alpha/2}(\nu)s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $\bar{x}_U = \bar{x}_d + t_{\alpha/2}(\nu)s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ 若 $d_0 \leq \bar{x}_L$ 或 $d_0 \geq \bar{x}_U$ ，则拒绝 H_0	$\bar{x}_d = \bar{x}_1 - \bar{x}_2$ $\bar{x}_U = \bar{x}_d + t_{\alpha}(\nu)s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ 若 $d_0 \geq \bar{x}_U$ ，则拒绝 H_0	$\bar{x}_d = \bar{x}_1 - \bar{x}_2$ $\bar{x}_L = \bar{x}_d - t_{\alpha}(\nu)s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ 若 $d_0 \leq \bar{x}_L$ ，则拒绝 H_0

例题 11.3 （见网络资源）

11.5 两个总体平均数检验，方差未知且不等

总体方差未知且不等，双总体平均数检验的假定条件（如图 11-3 所示）：

- (1) 两组样本是独立的，而且分别从两个总体随机抽样。
 - (2) 两个总体均为正态分布（或样本量大于等于 30）。
 - (3) 两总体的标准差未知且不等。
 - (4) n_1, n_2 为样本量， \bar{x}_1, \bar{x}_2 为样本平均数， s_1, s_2 为样本标准差。
- 决策法则如表 11-4 所示。

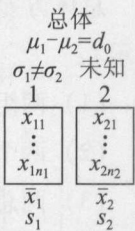


图 11-3 本节数据及假定条件

表 11-4 决策法则（两个总体平均数检验，方差未知且不等）

决策法则	双侧检验 $\begin{cases} H_0: \mu_1 - \mu_2 = d_0 \\ H_1: \mu_1 - \mu_2 \neq d_0 \end{cases}$	左侧检验 $\begin{cases} H_0: \mu_1 - \mu_2 \geq d_0 \\ H_1: \mu_1 - \mu_2 < d_0 \end{cases}$	右侧检验 $\begin{cases} H_0: \mu_1 - \mu_2 \leq d_0 \\ H_1: \mu_1 - \mu_2 > d_0 \end{cases}$
拒绝域法	$x_L = d_0 - t_{\alpha/2}(\nu) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $x_U = d_0 + t_{\alpha/2}(\nu) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$ <p>若 $\bar{x}_1 - \bar{x}_2 \leq x_L$ 或 $\bar{x}_1 - \bar{x}_2 \geq x_U$， 则拒绝 H_0</p>	$x_L = d_0 - t_{\alpha}(\nu) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$ <p>若 $\bar{x}_1 - \bar{x}_2 \leq x_L$，则拒绝 H_0</p>	$x_U = d_0 + t_{\alpha}(\nu) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$ <p>若 $\bar{x}_1 - \bar{x}_2 \geq x_U$，则拒绝 H_0</p>
检验值法	$t^* = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$ <p>若 $t^* \geq t_{\alpha/2}(\nu)$，则拒绝 H_0</p>	$t^* = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$ <p>若 $t^* \leq -t_{\alpha}(\nu)$，则拒绝 H_0</p>	$t^* = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$ <p>若 $t^* \geq t_{\alpha}(\nu)$，则拒绝 H_0</p>
p 值法	$p \text{ 值} = 2P[t(\nu) \geq t^*]$ <p>若 $p \text{ 值} < \alpha$，则拒绝 H_0</p>	$p \text{ 值} = P[t(\nu) \leq t^*]$ <p>若 $p \text{ 值} < \alpha$，则拒绝 H_0</p>	$p \text{ 值} = P[t(\nu) \geq t^*]$ <p>若 $p \text{ 值} < \alpha$，则拒绝 H_0</p>
置信区间法	$\bar{x}_L = \bar{x}_1 - \bar{x}_2 - t_{\alpha/2}(\nu) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $\bar{x}_U = \bar{x}_1 - \bar{x}_2 + t_{\alpha/2}(\nu) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ <p>若 $d_0 \leq \bar{x}_L$ 或 $d_0 \geq \bar{x}_U$， 则拒绝 H_0</p>	$\bar{x}_U = \bar{x}_1 - \bar{x}_2 + t_{\alpha}(\nu) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ <p>若 $d_0 \geq \bar{x}_U$，则拒绝 H_0</p>	$\bar{x}_L = \bar{x}_1 - \bar{x}_2 - t_{\alpha}(\nu) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ <p>若 $d_0 \leq \bar{x}_L$，则拒绝 H_0</p>

例题 11.4 （见网络资源）

11.6 两个总体平均数检验，样本是配对数据

样本是配对数据，双总体平均数检验的假定条件（如图 11-4 所示）：

- (1) 两组样本是“总体内”独立抽样，“总体间”配对抽样。
 $\{x_{1i}, i = 1, \dots, n\}$ 是独立抽样， x_{1i} 和 x_{2i} 是配对， $i = 1, \dots, n$ ，
 $d_i = x_{1i} - x_{2i}, i = 1, \dots, n$ ， $\bar{d} = \bar{x}_1 - \bar{x}_2$ 是 d_i 的平均数， s_d 是 d_i 的标准差。两组样本量 n 相等。
 - (2) 样本观察值之差 d_i 为正态分布（或样本量大于等于 30）。
 - (3) 两总体标准差未知。
- 决策法则如表 11-5 所示。

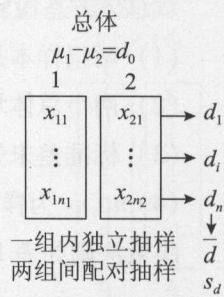


图 11-4 本节数据及假定条件

表 11-5 决策法则（两个总体平均数检验，样本是配对数据）

决策法则	双侧检验 $\begin{cases} H_0: \mu_1 - \mu_2 = d_0 \\ H_1: \mu_1 - \mu_2 \neq d_0 \end{cases}$	左侧检验 $\begin{cases} H_0: \mu_1 - \mu_2 \geq d_0 \\ H_1: \mu_1 - \mu_2 < d_0 \end{cases}$	右侧检验 $\begin{cases} H_0: \mu_1 - \mu_2 \leq d_0 \\ H_1: \mu_1 - \mu_2 > d_0 \end{cases}$
拒绝域法	$x_L = d_0 - t_{\alpha/2}(n-1) \frac{s_d}{\sqrt{n}}$ $x_U = d_0 + t_{\alpha/2}(n-1) \frac{s_d}{\sqrt{n}}$ 若 $\bar{x}_1 - \bar{x}_2 \leq x_L$ 或 $\bar{x}_1 - \bar{x}_2 \geq x_U$ ， 则拒绝 H_0	$x_L = d_0 - t_{\alpha}(n-1) \frac{s_d}{\sqrt{n}}$ 若 $\bar{x}_1 - \bar{x}_2 \leq x_L$ ，则拒绝 H_0	$x_U = d_0 + t_{\alpha}(n-1) \frac{s_d}{\sqrt{n}}$ 若 $\bar{x}_1 - \bar{x}_2 \geq x_U$ ，则拒绝 H_0
检验值法	$t^* = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{s_d / \sqrt{n}}$ 若 $ t^* \geq t_{\alpha/2}(n-1)$ ， 则拒绝 H_0	$t^* = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{s_d / \sqrt{n}}$ 若 $t^* \leq -t_{\alpha}(n-1)$ ， 则拒绝 H_0	$t^* = \frac{\bar{x}_1 - \bar{x}_2 - d_0}{s_d / \sqrt{n}}$ 若 $t^* \geq t_{\alpha}(n-1)$ ， 则拒绝 H_0
p 值法	p 值 = $2P[t(\nu) \geq t^*]$ 若 p 值 < α ，则拒绝 H_0	p 值 = $P[t(\nu) \leq t^*]$ 若 p 值 < α ，则拒绝 H_0	p 值 = $P[t(\nu) \geq t^*]$ 若 p 值 < α ，则拒绝 H_0
置信区间法	$\bar{x}_d = \bar{x}_1 - \bar{x}_2$ $\bar{x}_L = \bar{d} - t_{\alpha/2}(n-1) \frac{s_d}{\sqrt{n}}$ $\bar{x}_U = \bar{d} + t_{\alpha/2}(n-1) \frac{s_d}{\sqrt{n}}$ 若 $d_0 \leq \bar{x}_L$ 或 $d_0 \geq \bar{x}_U$ ，则拒绝 H_0	$\bar{x}_d = \bar{x}_1 - \bar{x}_2$ $\bar{x}_U = \bar{d} + t_{\alpha}(n-1) \frac{s_d}{\sqrt{n}}$ 若 $d_0 \geq \bar{x}_U$ ，则拒绝 H_0	$\bar{x}_d = \bar{x}_1 - \bar{x}_2$ $\bar{x}_L = \bar{d} - t_{\alpha}(n-1) \frac{s_d}{\sqrt{n}}$ 若 $d_0 \leq \bar{x}_L$ ，则拒绝 H_0

例题 11.5 （见网络资源）

例题 11.6 （见网络资源）

11.7 两个总体方差检验

双总体方差检验的假定条件（如图 11-5 所示）：

- (1) 两组样本是独立，而且分别从两个总体随机抽样。
- (2) 两个总体均为正态分布（或样本量大于等于 30）。
- (3) 标准差未知。
- (4) n_1, n_2 为样本量， s_1, s_2 为样本标准差。

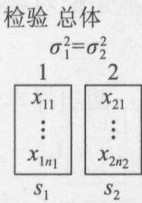


图 11-5 本节数据及假定条件

决策法则如表 11-6 所示。

表 11-6 决策法则（两个总体方差检验）

决策法则	双侧检验 $\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 \neq \sigma_2^2 \end{cases}$	左侧检验 $\begin{cases} H_0: \sigma_1^2 \geq \sigma_2^2 \\ H_1: \sigma_1^2 < \sigma_2^2 \end{cases}$	右侧检验 $\begin{cases} H_0: \sigma_1^2 \leq \sigma_2^2 \\ H_1: \sigma_1^2 > \sigma_2^2 \end{cases}$
拒绝域法	$s_L^2 = F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ $s_U^2 = F_{\alpha/2}(n_1 - 1, n_2 - 1)$ 若 $\frac{s_1^2}{s_2^2} \leq s_L^2$ 或 $\frac{s_1^2}{s_2^2} \geq s_U^2$ ， 则拒绝 H_0	$s_L^2 = F_{1-\alpha}(n_1 - 1, n_2 - 1)$ 若 $\frac{s_1^2}{s_2^2} \leq s_L^2$ ，则拒绝 H_0	$s_U^2 = F_{\alpha}(n_1 - 1, n_2 - 1)$ 若 $\frac{s_1^2}{s_2^2} \geq s_U^2$ ，则拒绝 H_0
检验值法	$f^* = \frac{s_1^2}{s_2^2}$ 若 $f^* \leq F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ ， 或 $f^* \geq F_{\alpha/2}(n_1 - 1, n_2 - 1)$ 则拒绝 H_0	$f^* = \frac{s_1^2}{s_2^2}$ 若 $f^* \leq F_{1-\alpha}(n_1 - 1, n_2 - 1)$ ， 则拒绝 H_0	$f^* = \frac{s_1^2}{s_2^2}$ 若 $f^* \geq F_{\alpha}(n_1 - 1, n_2 - 1)$ ， 则拒绝 H_0
p 值法	p 值 = $2\min\{P[F(n_1 - 1, n_2 - 1) \leq f^*], P[F(n_1 - 1, n_2 - 1) \geq f^*]\}$ 若 p 值 < α ，则拒绝 H_0	p 值 = $P[F(n_1 - 1, n_2 - 1) \leq f^*]$ 若 p 值 < α ，则拒绝 H_0	p 值 = $P[F(n_1 - 1, n_2 - 1) \geq f^*]$ 若 p 值 < α ，则拒绝 H_0
置信区间法	$\bar{s}_L^2 = \frac{s_1^2}{s_2^2} F_{1-\alpha/2}(n_2 - 1, n_1 - 1)$ $\bar{s}_U^2 = \frac{s_1^2}{s_2^2} F_{\alpha/2}(n_2 - 1, n_1 - 1)$ 若 $1 \leq \bar{s}_L^2$ 或 $1 \geq \bar{s}_U^2$ ，则拒绝 H_0	$\bar{s}_U^2 = \frac{s_1^2}{s_2^2} F_{\alpha}(n_2 - 1, n_1 - 1)$ 若 $1 \geq \bar{s}_U^2$ ，则拒绝 H_0	$\bar{s}_L^2 = \frac{s_1^2}{s_2^2} F_{1-\alpha}(n_2 - 1, n_1 - 1)$ 若 $1 \leq \bar{s}_L^2$ ，则拒绝 H_0

例题 11.7 （见网络资源）

11.8 两个总体比例检验

利用正态分布，双总体比例检验的假定条件（如图 11-6 所示）：

- (1) 两总体为贝努里分布， π_1, π_2 是两总体的成功率（比例）， d_0 是假设的两总体成功率之差。
- (2) 两组样本是独立，而且分别从两个总体随机抽样。
- (3) n_1, n_2 分别是两总体抽样的数目（样本容量）， t_1, t_2 分别是抽样的成功次数， $p_1 = t_1/n_1, p_2 = t_2/n_2$ 分别是抽样的成功率。
- (4) n_1, n_2 为大样本： $n_1 p_1 > 5, n_1 (1 - p_1) > 5$ ，且 $n_2 p_2 > 5, n_2 (1 - p_2) > 5$ 。

(5) $p_1 - p_2$ 近似正态分布，即 $p_1 - p_2 \sim N(d_0, \sigma_{p_1-p_2}^2)$ （决策法则是根据 H_0 成立）

- (6) 若 $d_0 = 0$ ，则

$$s_{p_1-p_2} = \sqrt{\left(\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}\right) \left(1 - \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}\right) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- (7) 若 $d_0 \neq 0$ ，则

$$s_{p_1-p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

决策法则如表 11-7 所示。

表 11-7 决策法则（利用正态分布，两个总体比例检验）

决策法则	双侧检验 $\begin{cases} H_0: \pi_1 - \pi_2 = d_0 \\ H_1: \pi_1 - \pi_2 \neq d_0 \end{cases}$	左侧检验 $\begin{cases} H_0: \pi_1 - \pi_2 \geq d_0 \\ H_1: \pi_1 - \pi_2 < d_0 \end{cases}$	右侧检验 $\begin{cases} H_0: \pi_1 - \pi_2 \leq d_0 \\ H_1: \pi_1 - \pi_2 > d_0 \end{cases}$
拒绝域法	$p_L = d_0 - z_{\alpha/2} s_{p_1=p_2}$ $p_U = d_0 + z_{\alpha/2} s_{p_1=p_2}$ 若 $p_1 - p_2 \leq p_L$ 或 $p_1 - p_2 \geq p_U$ ， 则拒绝 H_0	$p_L = d_0 - z_{\alpha} s_{p_1=p_2}$ 若 $p_1 - p_2 \leq p_L$ ，则拒绝 H_0	$p_U = d_0 + z_{\alpha} s_{p_1=p_2}$ 若 $p_1 - p_2 \geq p_U$ ，则拒绝 H_0
检验值法	$z^* = \frac{p_1 - p_2 - d_0}{s_{p_1=p_2}}$ 若 $ z^* \geq z_{\alpha/2}$ ，则拒绝 H_0	$z^* = \frac{p_1 - p_2 - d_0}{s_{p_1=p_2}}$ 若 $z^* \leq -z_{\alpha}$ ，则拒绝 H_0	$z^* = \frac{p_1 - p_2 - d_0}{s_{p_1=p_2}}$ 若 $z^* \geq z_{\alpha}$ ，则拒绝 H_0
p 值法	$p \text{ 值} = 2P(Z \geq z^*)$ 若 $p \text{ 值} < \alpha$ ，则拒绝 H_0	$p \text{ 值} = P(Z \leq z^*)$ 若 $p \text{ 值} < \alpha$ ，则拒绝 H_0	$p \text{ 值} = P(Z \geq z^*)$ 若 $p \text{ 值} < \alpha$ ，则拒绝 H_0

检验 总体

$\pi_1 - \pi_2 = d_0$	
1	2
1	0
0	1
1	1
\vdots	\vdots
0	1
0	0
1	0
Σ	$\begin{matrix} t_1 & t_2 \\ n_1 & n_2 \\ p_1 & p_2 \end{matrix}$

图 11-6 本节数据及假定条件

请注意：当 $d_0 = 0$ 或 $d_0 \neq 0$ ，计算置信区间是不相同的，因为决策法则是根据 H_0 成

立，但是在 11.2.2 节计算置信区间时不考虑 $d_0 = 0$ ，为了避免混淆，不列置信区间法。

例题 11.8 (见网络资源)

11.9 中文统计应用

11.9.1 两个总体 t 检验 (例题 11.6)

执行 t 检验的操作示意图和结果如图 11-7 所示。

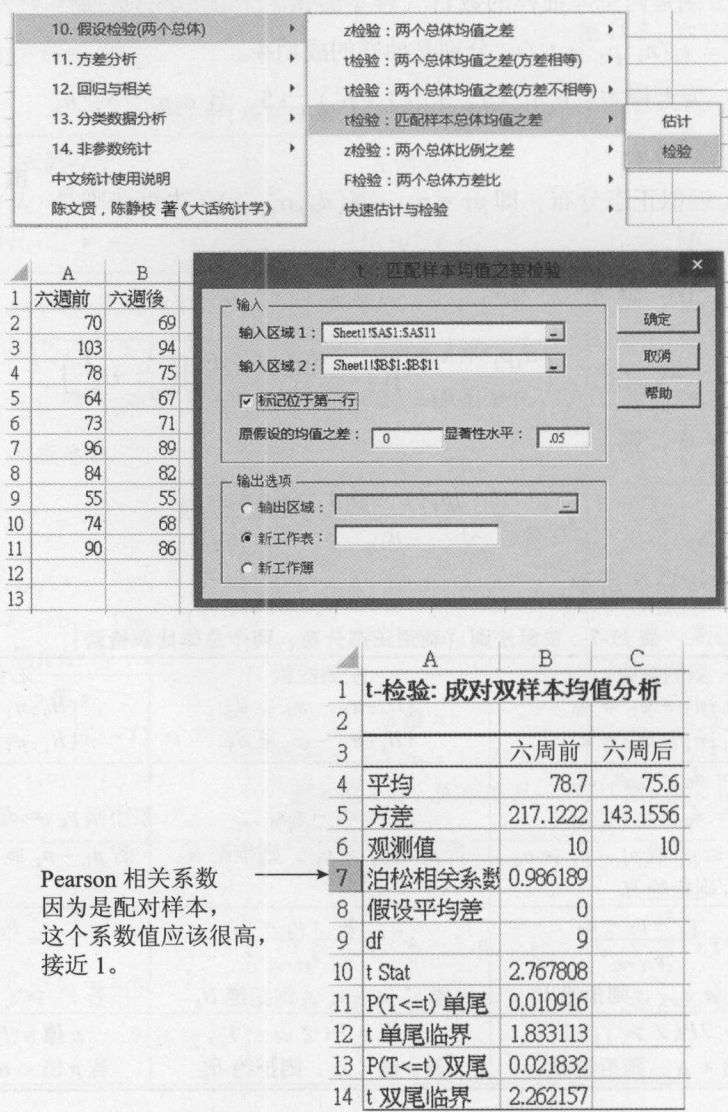


图 11-7 执行 t 检验的操作示意图和结果

11.9.2 快速估计与检验（例题 11.7，例题 11.8）

执行“快速估计与检验”的示例如图 11-8 所示。

比例值差检验(假设比例值差等于 0)			比例值差检验(假设比例值差不等于 0)			F 检验-双总体方差检定	
平均数差检定, 变异数已知			平均数差检定, 变异数未知但相等			平均数差检定, 变异数未知且不等	
	A	B	C	D	E	F	
1	z 检定: 两个总体比例检验(假设检验比例值差不等于 0)						
2							
3	输入数据:						
4							
5		样本1	样本2				
6	样本比例	0.0600	0.0300				
7	样本量	100	120				
8	假设的比例之差	0.1	(不等于 0)				
9	显著性水平(α)	0.05			重新计算		
10							
11	输出数据:						
12							
13	z 值	-2.4649					
14	P(Z>= z) 单侧p值	0.0069					
15	临界值: 右侧	1.6449					
16	临界值: 左侧	-1.6449					
17	2P(Z>= z) 双侧p值	0.0137					
18	临界值: 双侧	1.9600					

	A	B	C	D	E	F	G
1	F 检验: 两个总体方差检定						
2							
3	输入数据:						
4							
5		样本1	样本2				
6	样本方差	0.0324	0.0100				
7	样本量	21	25				
8	显著性水平(α)	0.05					
9							
10	输出数据:(利用表9.6)		输出数据:(利用表9.7)				
11							
12	F 值(F^*)	3.2400		双侧检定	左侧检定	右侧检定	
13	P(F>= F^*) 单侧p值	0.0035		H0: $\sigma_1 = \sigma_2$	H0: $\sigma_1 \geq \sigma_2$	H0: $\sigma_1 \leq \sigma_2$	
14	临界值: 右侧	2.0267		F 值(F^*)	3.2400	0.3086	3.2400
15	临界值: 左侧	0.4802		临界值	2.3273	2.0825	2.0267
16	2P(F>= F^*) 双侧p值	0.0069		p 值	0.0069	0.9965	0.0035
17	临界值: 双侧	0.4154		决策法则	若 F 值(F^*) > 临界值, 则拒绝 H0		
18		2.3273					
19							
20		重新计算					
21							

图 11-8 执行“快速估计与检验”的示例

两个总体参数(均值, 比例, 方差, 中位数)检验 总体数 $k=2$									
定距尺度					定类尺度 两个总体比例				
两个总体均值		两个总体方差			定距尺度		独立样本 配对样本		
独立样本		配对样本			独立样本		配对样本		
非正态总体 $n < 30$		正态总体			非正态总体 $n < 30$		正态总体		
正总体标准差 σ_1 已知		正总体标准差 σ_1 未知			非正态总体 $n < 30$		正态总体		
$\sigma_1 = \sigma_2$		$\sigma_1 \neq \sigma_2$			非正态总体 $n < 30$		正态总体		
秩总和检验 15.5节	z检验 $\mu_1 - \mu_2$ 11.3节	t检验 $\mu_1 - \mu_2$ 11.4节	t检验 $\mu_1 - \mu_2$ 11.5节	t检验 μ_D 11.6节	符号秩检验 15.3节	F检验 σ_1^2 / σ_2^2 11.7节	秩总和检验 15.5节	符号检验 15.3节	超几何检验 15.12节
$k > 2$	$k > 2$	$k > 2$	$k > 2$	$k > 2$	$k > 2$	$k > 2$	$k > 2$	$k > 2$	$k > 2$
KW 检验 15.6节	单因素 ANOVA 12.2节, 12.3节		随机区组设计 12.7节	Friedman 检验 15.7节	Bartlett 检验 12.5节	KW 检验 15.6节	Friedman 检验 15.7节	卡方检验 14.10节	置信区间 11.2节 McNemar 14.11节
									($k > 2$ 两个以上总体)

图 11-9 第 11 章流程图

11.11 本章思维导图

本章思维导图如图 11-10 所示。

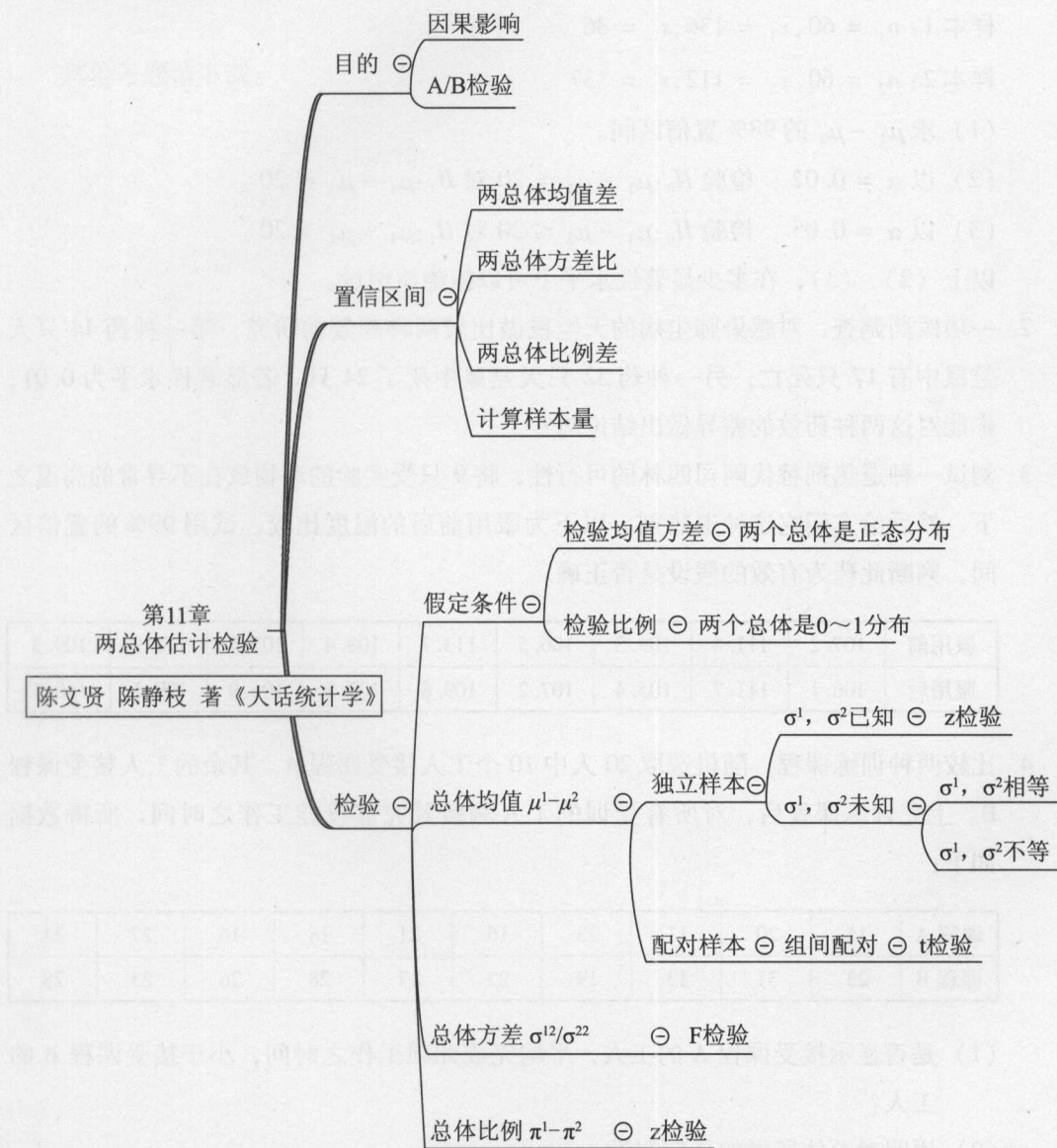


图 11-10 第 11 章思维导图

习 题

1. 下面的统计量数为取自两个不同总体的独立样本：
- 样本 1: $n_1 = 60, \bar{x}_1 = 136, s_1^2 = 86$
- 样本 2: $n_2 = 60, \bar{x}_2 = 112, s_2^2 = 137$
- (1) 求 $\mu_1 - \mu_2$ 的 98% 置信区间。
- (2) 以 $\alpha = 0.02$ ，检验 $H_0: \mu_1 - \mu_2 = 20$ 对 $H_1: \mu_1 - \mu_2 \neq 20$ 。
- (3) 以 $\alpha = 0.05$ ，检验 $H_0: \mu_1 - \mu_2 \leq 20$ 对 $H_1: \mu_1 - \mu_2 > 20$ 。
- 以上 (2)，(3)，在多少显著性水平下可以拒绝原假设。
2. 一项医药调查，对感染肺尘病的天竺鼠做比较两种药效的研究，第一种药 44 只天竺鼠中有 17 只死亡；另一种药 52 只天竺鼠中死了 24 只。若显著性水平为 0.01，你能对这两种药效的差异做出结论吗？
3. 测试一种退热剂替代阿司匹林的可行性，将 9 只受实验的动物放在不寻常的高温之下，然后给它们吃这种退热剂。以下为服用前后的温度比较，试用 99% 的置信区间，判断此药为有效的假设是否正确。

服用前	107.2	111.4	109.3	106.5	113.7	108.4	107.7	111.9	109.3
服用后	106.1	111.7	105.4	107.2	109.8	108.8	106.9	109.6	110.5

4. 比较两种训练课程，随机选取 20 人中 10 个工人接受课程 A，其余的工人接受课程 B。上完训练课程后，对所有受训的工人测验其完成技能工作之时间，所得数据如下：

课程 A	15	20	11	23	16	21	18	16	27	24
课程 B	23	31	13	19	23	17	28	26	25	28

- (1) 是否显示接受课程 A 的工人，平均完成此项工作之时间，小于接受课程 B 的工人？
- (2) 说明对总体所做的任何假设。
- (3) 两课程之完成工作时间的总体的平均差之 95% 置信区间。
5. 某大学商学院的学生必修统计课，该学院开出两班统计课，从每班修课学生中各随机抽样 40 名，得到以下的期末分数：第一班样本平均数为 76 分、标准差为 8

分；第二班样本平均数为 71 分、标准差为 7.5 分，假设期末分数为正态分配（总分为 100 分），请根据这两个样本回答以下的问题。

- (1) 这两班的统计课期末分数是否相同？请用 5% 的显著性水平检验。
- (2) 请计算两班统计课期末分数差检验之 p 值。
- (3) 请用 95% 的置信区间，检验这两班的统计课期末分数是否相同？

其他习题请下载。



第12章

方差分析

于诸果中、应说何果，何因所得。

——佛经《俱舍论》（卷六）

是仁义用于古，不用于今也，故曰：“世异则事异”。

是干戚用于古，不用于今也，故曰：“事异则备变”。

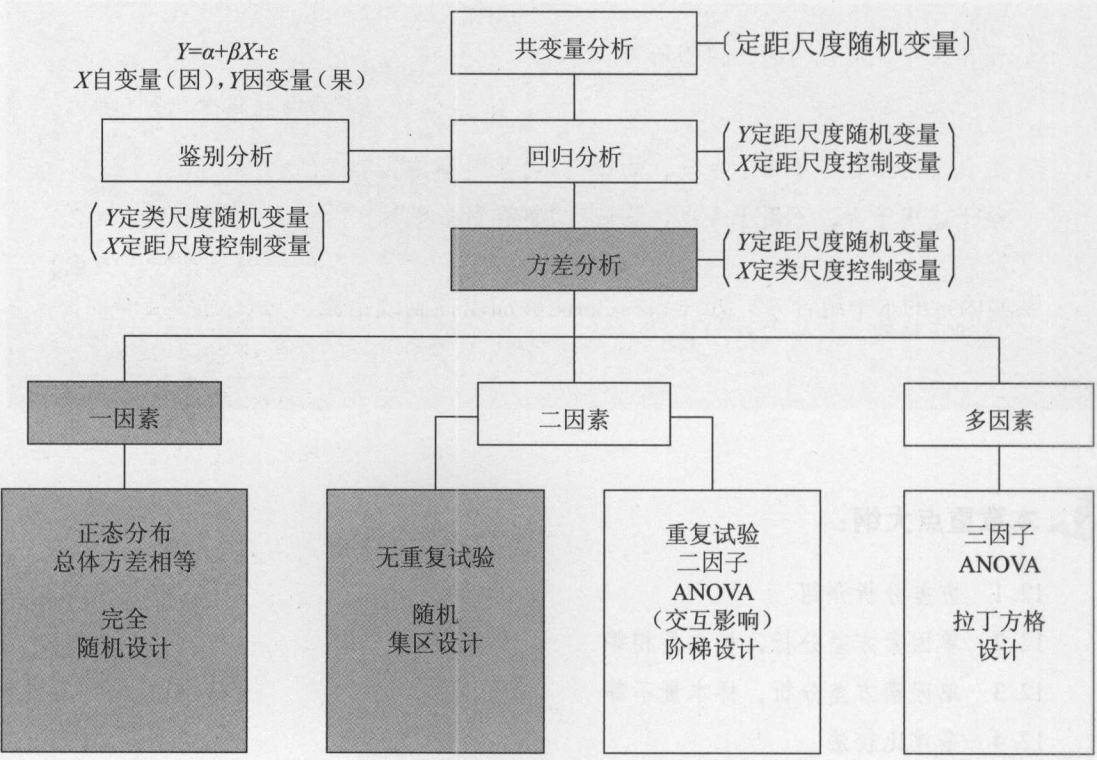
——《韩非子·五蠹》

致知在格物，物格而后知至。

——《礼记·大学》

本章重点大纲：

- 12.1 方差分析介绍
- 12.2 单因素方差分析，样本量相等
- 12.3 单因素方差分析，样本量不等
- 12.4 多重比较法
- 12.5 检验方差是否相等
- 12.6 参数估计
- 12.7 双因素方差分析，无交互作用
- 12.8 中文统计应用
- 12.9 本章流程图
- 12.10 本章思维导图



本章概念图 (浅蓝色为本章内容)

12.1 方差分析介绍

方差分析 (analysis of variance, 简称 ANOVA) 是计算方差, 用以比较两个以上总体的“平均数”是否全部相等。方差分析其实是前一章, “两个总体平均数检验, 方差未知但相等”, 扩展到两个以上的总体。

方差分析, 是检验同一因素或多因素的数种处理 (可视为数个总体) 的平均数, 是否相等; 或者不同因素之间是否有交互影响。

因素 (factor) 是质量标志, 是定类变量或定序变量, 其分类的标志值, 称为水平 (level)。

因素有实验因素 (experiment factor) 与观察因素 (observation factor), 每个因素有不同的水平, 不同因素的特定水平组合, 为一个处理 (treatment) 或区组 (block)。

实验因素的水平组合为“处理”, 我们能够加以控制或指派, 可以随意给样本任何水平或处理。例如: 某个样本 (某个人或动物) 可以用 A 种药物, 也可以用 B 种药物; 此外, 实验的“方式” (对产品产量的影响), 样本的“颜色” (对销售量的影响), 班级的“人数” (对升学率的影响), 教学的“老师” (对教学效果的影响), 可以算是实验因素。

观察因素的水平为“区组”, 我们不能加以控制, 不能随意给样本任何其他区组。例如: 样本的“血型”, 某个样本 (某个人) 的血型是 A 型, 我们不能指定他是 B 型。此外, 非温室实验的“天气” (对产品产量的影响), 样本的“性向” (对学生成绩的影响), 家庭的“人数” (对家庭总收入的影响), 可以算是观察因素。例如: 同一个工人 (区组) 操作 3 台机器 (处理) 的时间或产量 (观测值)。

方差分析 ANOVA 有: 单因素 ANOVA, 双因素 ANOVA, 三因素 ANOVA 等。ANOVA 和实验设计有密切的关系。ANOVA 也是因果关系, “因”是因素, “果”是反应变量 (response variable) 观测值, 或者观测值的平均数。

单因素 ANOVA 的实验设计是完全随机设计 (complete randomized design), 假设单因素有 k 个水平即 k 个处理, 每个处理有 n_i 个样本, 总共有 $N = n_1 + n_2 + \cdots + n_k$ 个样本, 然后随机选 n_1 个样本指定第一个处理, 随机选 n_2 个样本指定第二个处理, 依此类推, 最后随机选 n_k 个样本指定第 k 个处理, 然后记录其反应变量观测值。当处理数目等于 2 时, “完全随机设计”相当于“双总体独立样本平均数检验 (方差未知但相等)”。

随机化区组设计 (randomized block design) 是两个因素, 一个是实验因素, 一个是观察因素。观察因素的区组数目是 a , 实验因素的处理数目是 k 。当处理数目等于 2 时, “随

机化区组设计”相当于“双总体匹配样本平均数检验”。因为双总体匹配样本可以消除一些外在因素，所以随机化区组设计也可以消除一些外在因素。因为同一区组代表相关的样本，例如同一个人经历两种处理。所以“随机化区组设计”与“双总体匹配样本”一样，可以将一些外在的变量消除掉。

单因素方差分析，因素的每个水平或处理，是一个总体。多因素方差分析，不同因素的水平组合（甲因素水平 A 加乙因素水平 B），是一个总体。

方差分析是检验这些总体的“均值”，是否有显著的差异？因素的影响是否显著？

方差分析 ANOVA 的假定条件如下。

- (1) 总体反应变量的观测值是正态分布或近似正态分布。
- (2) 样本是独立抽样（各总体之内样本独立，总体之间样本也是独立）。
- (3) 总体反应变量的标准差未知但相等。

如图 12-1 所示。

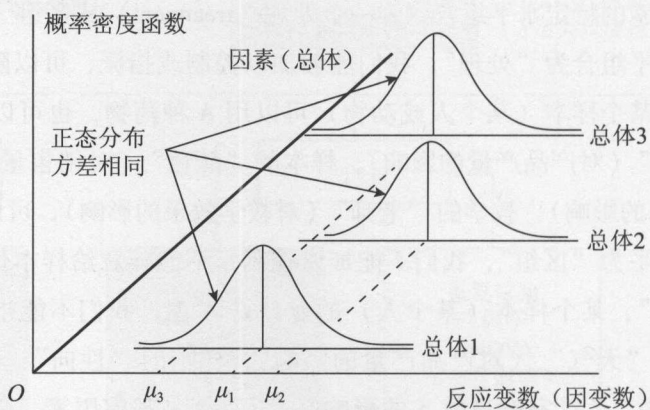


图 12-1 方差分析之假定条件

方差分析是根据“方差”（变异的总和）的分解，检验总体的“均值”。方差分析主要是根据总体之间（组间）的变异和总体之内（组内）的变异，检验总体均值是否相等。如果组间变异大，组内变异小，则总体均值有显著不相等。

如图 12-2 所示，有 3 个总体。如果 3 个总体的组内变异大，如图 12-2（a），组间变异与图 12-2（b）相同，则方差分析检验结果：3 个总体的均值，没有显著不相等。

如果 3 个总体的组内变异小，组间变异和图 12-2（a）相同，则方差分析检验结果：3 个总体的均值有显著不相等。

以上只是概念图示，我们当然要计算统计量 F ， $F = \text{组间方差} / \text{组内方差}$ ， F 越大，显著性越高。

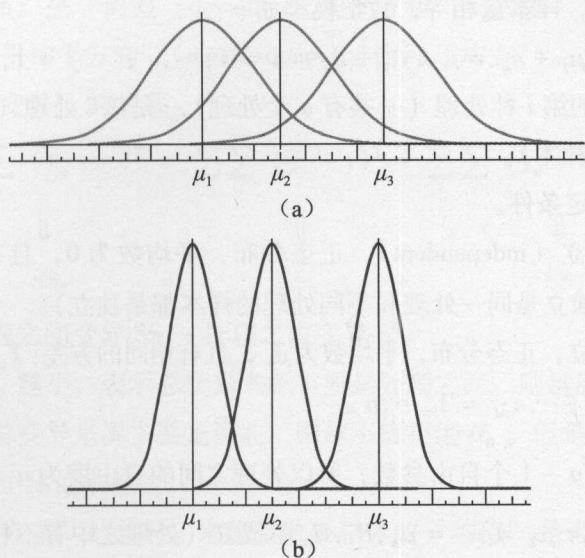


图 12-2 方差分析是根据方差（变异），检验多个总体的均值

12.2 单因素方差分析，样本量相等

单因素 ANOVA 的实验设计是完全随机设计（complete randomized design），如表 12-1 所示。

表 12-1 单因素 ANOVA 的实验设计

总体（实验因素/处理）			
1	2	...	a
Y_{11}	Y_{21}	...	Y_{a1}
\vdots	\vdots	\vdots	\vdots
Y_{1n}	Y_{2n}	...	Y_{an}

有关符号及名词定义说明如下：

- Y_{ij}, y_{ij} ——第 i 个处理的第 j 个样本的反应变量及观测值；
- μ_i ——第 i 个处理的总体均值（未知参数）；
- α_i ——第 i 个处理的效应（未知参数）；
- σ^2 ——每个处理的相同方差（未知参数）；
- μ ——所有处理的总均值（未知参数）；
- ε_{ij} ——第 i 个处理的第 j 个样本的误差（随机变量）。

单因素方差分析, 样本量相等, 数学模型如下:

$$Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \dots, a; \quad j = 1, \dots, n$$

上述 i 是因素中的第 i 种处理 (一共有 a 个处理), j 是该 i 处理 (该总体) 的第 j 个样本 (每个处理有 n 个样本)。

下列是模型的假定条件。

(1) ε_{ij} 为互相独立 (independent), 正态分布, 平均数为 0, 且有相同的方差: $\varepsilon_{ij} \sim N(0, \sigma^2) \forall i, j$ (互相独立是同一处理和不同处理的样本都是独立)。

(2) Y_{ij} 为互相独立, 正态分布, 平均数为 μ_i , 且有相同的方差: $Y_{ij} \sim N(\mu_i, \sigma^2)$ 或 $Y_{ij} \sim N(\mu + \alpha_i, \sigma^2) \forall i = 1, \dots, a; j = 1, \dots, n$ 。

(3) $\sum_{i=1}^a \alpha_i = 0$ ($a - 1$ 个自由参数, 所以处理之间的自由度为 $a - 1$)。

(4) 检验 $H_0: \mu_1 = \mu_2 = \dots = \mu_a, H_1: H_0$ 不成立 (处理之中有不相等的平均数)。

或者检验 $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_a = 0, H_1: H_0$ 不成立 (处理因素效应有显著)。

要进行方差分析, 定义符号及名词:

\bar{y}_i : 第 i 个处理的样本观测值之平均数, 即

$$\bar{y}_i = \sum_{j=1}^n y_{ij} / n$$

\bar{y} : 所有处理的样本观测值之总平均数, 即

$$\bar{y} = \sum_{i=1}^a \sum_{j=1}^n y_{ij} / an$$

差异: 观测值与总平均数之差异, 即

$$\text{差异} = y_{ij} - \bar{y}$$

处理间差异: 处理的平均数与总平均数之差异, 即

$$\text{处理间差异} = \bar{y}_i - \bar{y}$$

处理内差异: 观测值与所属处理的平均数之差异, 这是随机误差或称残差, 即

$$\text{处理内差异} = y_{ij} - \bar{y}_i$$

SS_T : “总变异”, 是总差异平方和, 表示不同处理之间的变异, 即

$$SS_T = \sum_{i,j} (y_{ij} - \bar{y})^2$$

SS_A : “组间变异”, 是处理间差异平方和, 表示不同处理之间的变异, 即

$$SS_A = \sum_{i,j} (\bar{y}_i - \bar{y})^2$$

SS_E : “组内变异”, 是处理内差异平方和, 残差平方和或误差平方和, 即

$$SS_E = \sum_{i,j} (y_{ij} - \bar{y}_i)^2$$

“变异” (variation) 是“离差 (deviation) 的平方和”，“方差” (variance) 是“变异除以自由度”或称“平均平方和” (mean square) (简称“均方”)。

下列公式说明“变异”的关系：

$$\begin{array}{ccccc} \sum_i \sum_j (y_{ij} - \bar{y})^2 & = & \sum_i \sum_j (\bar{y}_i - \bar{y})^2 & + & \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 \\ \Downarrow & & \Downarrow & & \Downarrow \\ SS_T & = & SS_A & + & SS_E \end{array}$$

总变异 SS_T = 处理之间变异 SS_A + 处理之内变异 SS_E

当 SS_A 越大, SS_E 越小, 表示总变异来源主要是处理之间, 则越能拒绝 H_0 。当 SS_A 越小, SS_E 越大, 表示总变异来源主要是误差, 则越不能拒绝 H_0 。但是要大到什么程度或小到什么程度, 需要一个比值来检验。

方差分析是以“方差”的差异, 来检验数个总体的“均值”是否相等, 即

$$MS_A = \frac{SS_A}{a-1} = \frac{n \sum_{i=1}^a (\bar{y}_i - \bar{y})^2}{a-1}$$

$$MS_E = \frac{SS_E}{a(n-1)} = \frac{\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2}{a(n-1)}$$

式中: MS_A , MS_E —— σ^2 的不偏估计量。

如果上述公式 y_{ij} 改为 Y_{ij} , 则 MS_A , MS_E 是随机变量。

MS_A 会高估 σ^2 , 但是在原假设 H_0 成立下 ($\alpha_i = 0$), $E(MS_A) = \sigma^2$, 则

$$\frac{(a-1)MS_A}{\sigma^2} \sim \chi^2(a-1), \frac{(an-a)MS_E}{\sigma^2} \sim \chi^2(an-a)$$

$$\Rightarrow F = \frac{MS_A}{MS_E} = \frac{\frac{(a-1)MS_A/\sigma^2}{a-1}}{\frac{(an-a)MS_E/\sigma^2}{an-a}} \sim F(a-1, an-a)$$

所以 ANOVA 是利用 F 检验, 表 12-2 是单因素方差分析, 样本量相等, 方差分析表。

表 12-2 方差分析表 (单因素方差分析, 样本量相等)

变异来源	平方和	自由度	均方	F 比值
处理之间 (因素)	$SS_A = n \sum_{i=1}^a (\bar{y}_i - \bar{y})^2$	$a-1$	$MS_A = \frac{SS_A}{a-1}$	$F = \frac{MS_A}{MS_E}$
处理之内 (误差)	$SS_E = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2$	$a(n-1)$	$MS_E = \frac{SS_E}{a(n-1)}$	
总和	$SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y})^2$	$an-1$		

若 $F \geq F_{\alpha}[a-1, a(n-1)]$, 则拒绝 H_0 。

注意:

(1) 如果已知各处理的样本方差 s_i^2 , 则 $SS_E = \sum_{i=1}^a (n-1)s_i^2$ 。

合并方差 s_p^2 是 σ^2 的估计值, 即 $s_p^2 = \frac{\sum (n-1)s_i^2}{an-a} = MS_E = \hat{\sigma}^2$

(2) 若已知各处理的样本均值 \bar{y}_i 和样本方差 s_i^2 , 则可用“中文统计”——“方差分析”——“快速检验”。

(3) ANOVA 表没有 MS_T , 即使有, 但是 $MS_A + MS_E \neq MS_T$ 。

例题 12.1 为了比较 4 种中文输入法 (编号: 甲, 乙, 丙, 丁) 的输入速度。每种输入法, 随机选出 5 个人, 输入同一篇文章。以下是两次实验的输入时间观测值 (分钟)。

实验设计 A

甲: 17, 24, 39, 42, 43

乙: 28, 32, 44, 50, 61

丙: 41, 45, 48, 54, 57

丁: 22, 29, 30, 34, 40

实验设计 B

甲: 2, 24, 39, 42, 58

乙: 13, 32, 44, 50, 76

丙: 26, 45, 48, 54, 72

丁: 7, 29, 30, 34, 55

检验这 4 种输入法的输入时间没有差异。显著性水平是 0.05, 问检验的结果如何?

解答: 实验设计 A 和 B 的数据分布如图 12-3 所示。令 $\mu_1, \mu_2, \mu_3, \mu_4$ 是甲、乙、丙、丁 4 种输入法的平均时间。

检验 $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$, $H_1: H_0$ 不成立。

计算, 得: $\bar{y}_1 = 33$, $\bar{y}_2 = 43$, $\bar{y}_3 = 49$, $\bar{y}_4 = 31$, $\bar{\bar{y}} = 39$ 。

实验设计 A 和 B 的处理甲、乙、丙、丁 (以 1, 2, 3, 4 为代号) 各组平均和总平均都相等, 但是方差不相等, 即

$$\bar{y}_1 = 33, \bar{y}_2 = 43, \bar{y}_3 = 49, \bar{y}_4 = 31, \bar{\bar{y}} = 39$$

所以实验设计 A 和 B 的 SS_A 相同, 即

$$SS_A = n \sum_{i=1}^a (\bar{y}_i - \bar{\bar{y}})^2 = 5 \times [(33 - 39)^2 + (43 - 39)^2 + (49 - 39)^2 + (31 - 39)^2] = 1080$$

实验设计 A 的各组方差

$$s_1^2 = 138.5, s_2^2 = 180, s_3^2 = 42.5, s_4^2 = 44$$

所以

$$SS_E = \sum_{i=1}^4 (n-1)s_i^2 = 4 \times (138.5 + 180 + 42.5 + 44) = 1620$$

实验设计 B 的各组方差

$s_1^2 = 446, s_2^2 = 540, s_3^2 = 275, s_4^2 = 291.5$

所以

$$SS_E = \sum_{i=1}^4 (n-1)s_i^2 = 4 \times (446 + 540 + 275 + 291.5) = 6210$$

实验设计 A 的方差分析表如表 12-3 所示。

表 12-3 实验设计 A 的方差分析表

变异来源	平方和	自由度	均方	F 比值
组间	$SS_A = 1080$	3	$MS_A = 360$	$F = 3.56$
组内	$SS_E = 1620$	16	$MS_E = 101.25$	
总和	$SS_T = 2700$	19		

因为 $F = 3.56 \geq F_{0.05}(3,16) = 3.239$ ，所以拒绝 H_0 。

实验设计 B 的方差分析表如表 12-4 所示。

表 12-4 实验设计 B 的方差分析表

变异来源	平方和	自由度	均方	F 比值
组间	$SS_A = 1080$	3	$MS_A = 360$	$F = 0.9275$
组内	$SS_E = 6210$	16	$MS_E = 388.125$	
总和	$SS_T = 7290$	19		

因为 $F = 0.9275 \leq F_{0.05}(3,16) = 3.239$ ，所以无法拒绝 H_0 。

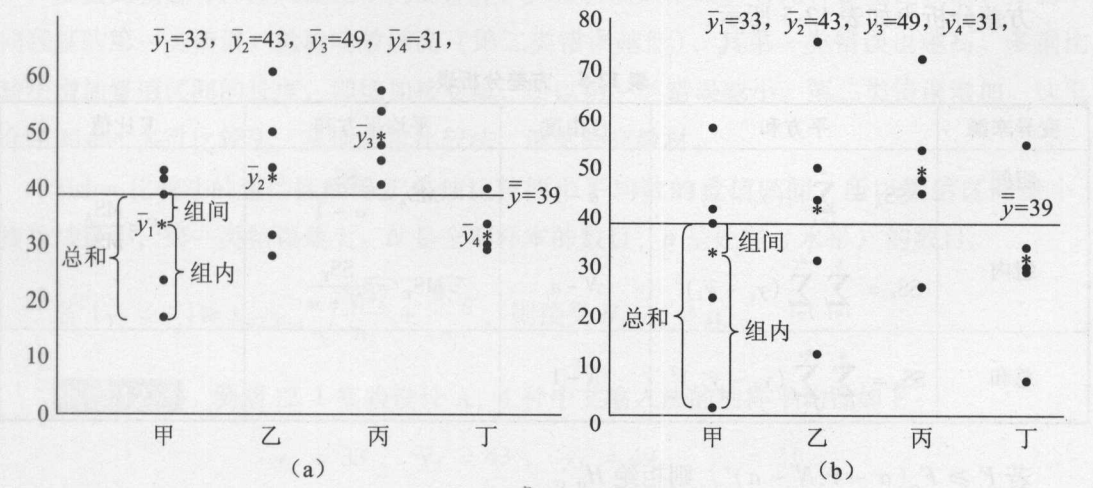


图 12-3 实验设计 A 和 B 的数据分布

(a) 实验设计 A 的数据分布；(b) 实验设计 B 的数据分布

实验设计 A 和 B 的处理甲、乙、丙、丁各组平均和总平均都相等，所以组间变异 SS_A 相等，但是实验设计 A 的各组方差较小，所以组内变异 SS_E 小，因此结论是甲、乙、丙、丁各组均值存在差异。实验设计 B 的各组方差较大，所以组内变异 SS_E 大，因此结论是甲、乙、丙、丁各组均值不存在差异。

12.3 单因素方差分析，样本量不等

单因素方差分析，样本量不等，数学模型如下：

$$Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \cdots, a; \quad j = 1, \cdots, n_i$$

上述 i 是因素中的第 i 种处理（一共有 a 个处理）， j 是该 i 处理（该总体）的第 j 个抽样（第 i 个处理有 n_i 个抽样）。

下列是模型的假定条件：

(1) ε_{ij} 为互相独立，正态分布，平均数为 0，且有相同的方差， $\varepsilon_{ij} \sim N(0, \sigma^2)$ 。

(2) Y_{ij} 为互相独立，正态分布，平均数为 μ_i ，且有相同的方差， $Y_{ij} \sim N(\mu_i, \sigma^2)$ 。

$$(3) \sum_{i=1}^a \alpha_i = 0。$$

(4) 检验 $H_0: \mu_1 = \mu_2 = \cdots = \mu_a, H_1: H_0$ 不成立。（处理之中有不相等的平均数）

方差分析表如表 12-5 所示。

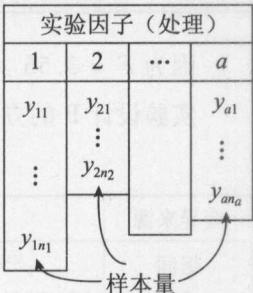


图 12-4 实验模型

表 12-5 方差分析表

变异来源	平方和	自由度	平均平方和	F 比值
组间	$SS_A = \sum_{i=1}^a n_i (\bar{y}_i - \bar{\bar{y}})^2$	$a - 1$	$MS_A = \frac{SS_A}{a - 1}$	$F = \frac{MS_A}{MS_E}$
组内	$SS_E = \sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$	$N - a$	$MS_E = \frac{SS_E}{N - a}$	
总和	$SS_T = \sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$	$N - 1$		

若 $F \geq F_\alpha(a - 1, N - a)$ ，则拒绝 H_0 。

注意：(1) 如果已知各处理的样本方差 s_i^2 ，则 $SS_E = \sum_{i=1}^a (n_i - 1)s_i^2$ 。

(2) 合并方差 s_p^2 是 σ^2 的估计值, 即

$$s_p^2 = \frac{\sum (n_i - 1)s_i^2}{N - a} = MS_E = \hat{\sigma}^2$$

例题 12.2—12.3 (见网络资源)

12.4 多重比较法

多重比较法 (multiple comparison) 是用在方差分析以后, 发现平均数有显著差异, 再找出每两个平均数做比较, 找出平均数大小的顺序。

多重比较法是“每对”平均数做比较, 利用置信区间, 检验每对平均数是否相等。

如果有 3 个总体 (处理), 其平均数分别为 μ_1, μ_2, μ_3 , 利用双总体 t 检验, 分别比较 3 对平均数 $\mu_1 - \mu_2, \mu_1 - \mu_3, \mu_2 - \mu_3$, 分别有第一类错误 α 。如果 3 个 t 检验是独立 (事实上不是, 因为检验中的样本数据有重复, 例如总体 1 的样本数据用在检验 $\mu_1 - \mu_2$ 及 $\mu_1 - \mu_3$), 则 3 个 t 检验中至少有一次第一类错误的概率是

$$1 - P(3 \text{ 个都没有第一类错误}) = 1 - (0.95)^3 = 1 - 0.86 = 0.14$$

3 个 t 检验至少有一个拒绝 $H_0: \mu_i - \mu_j$ 的第一类错误有 0.14。

拒绝 $H_0: \mu_1 = \mu_2 = \cdots = \mu_n \Leftrightarrow$ 至少否定一个 $H_0: \mu_i - \mu_j = 0$

所以 3 组 $\mu_i - \mu_j$ 分别用 t 检验, 则第一类错误提高。

多重比较法有许多人提出不同的计算, 其检验功效不同。有几种多重比较法, 控制不同程度的第一类错误。检验功效越高 (第二类错误越低), 其第一类错误也越高。多重比较法增加置信区间的长度, 即增加接受域, 所以第一类错误减小, 第二类错误增加。这里介绍 Fisher 多重比较法, 其他多重比较法, 请见补充教材。

Fisher 比较法的置信区间等于单独比较两个平均数的置信区间, 所以置信区间最小, 接受域最小, 第一类错误最大。N 是全部样本的数目, a 是处理 (水平) 的数目。

若 $|\bar{y}_i - \bar{y}_j| \geq t_{\alpha/2, N-a} \sqrt{\frac{MS_E}{n_i} + \frac{MS_E}{n_j}}$, 则接受 $H_1: \mu_i \neq \mu_j$ 。

例题 12.4 例题 12.1 实验设计 A, 4 种中文输入法的抽样平均数如下:

$$\bar{y}_1 = 33, \bar{y}_2 = 43, \bar{y}_3 = 49, \bar{y}_4 = 31$$

检验下列 6 ($C_2^4 = 6$) 个假设:

$$H_0^1: \mu_1 = \mu_2 \quad H_0^2: \mu_1 = \mu_3 \quad H_0^3: \mu_1 = \mu_4$$

$$H_0^4: \mu_2 = \mu_3 \quad H_0^5: \mu_2 = \mu_4 \quad H_0^6: \mu_3 = \mu_4$$

$$\text{解答: } t_{\frac{\alpha}{2}, a(n-1)} \sqrt{\frac{2MS_E}{n}} = t_{0.025, 16} \sqrt{\frac{2(101.25)}{5}} = 13.5$$

4 种中文输入法的抽样平均数, 由小到大排列如下:

		\bar{y}_1	\bar{y}_2	\bar{y}_3
		33	43	49
\bar{y}_4	31	2	12	18*
\bar{y}_1	33		10	16*
\bar{y}_2	43			6

上述*表示有显著差异 (>13.5)。即拒绝: $H_0^2: \mu_1 = \mu_3$ 和 $H_0^6: \mu_3 = \mu_4$ 。

12.5 检验方差是否相等

方差分析的假定条件之一是各总体 (处理) 的方差相等。F 检验对这个条件是稳健的 (robust), 所谓稳健的模型是如果假定条件 (assumption) 不符, 但是解答还可以适用于原来问题, ANOVA 的总体方差稍有不等, 而检验结果仍然可信。尤其是当样本量相等时, 方差分析检验更是稳健的。不过如果害怕方差相差很大, 就应该对方差是否相等做检验:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

检验的方法有几种, 我们介绍 Bartlett 检验法。

从第 i 个总体随机抽样 n_i 个样本。总样本量 $N = \sum_{i=1}^k n_i$ 。

计算每个总体的样本方差 s_i^2 , $i = 1, \dots, k$

$$s_i^2 = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n_i - 1}$$

$$s^2 = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{N - k} = \frac{\sum_{i=1}^k (n_i - 1) s_i^2}{N - k}$$

$$C = 1 + \frac{1}{3(k-1)} \left[\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{N - k} \right]$$

$$B = \frac{1}{C} \left[(N - k) \ln_e s^2 - \sum_{i=1}^k (n_i - 1) \ln_e s_i^2 \right]$$

$$B = \frac{\ln_e 10}{C} \left[(N - k) \lg_{10} s^2 - \sum_{i=1}^k (n_i - 1) \lg_{10} s_i^2 \right] \qquad \ln_e 10 = 2.3026$$

当 $B > \chi^2_{\alpha, k-1}$, 则拒绝 H_0 。

例题 12.5 4 种中文输入法的抽样方差: $s_1^2 = 138.5, s_2^2 = 180, s_3^2 = 42.5, s_4^2 = 44$ 。

检验: $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$, 如果显著性水平是 0.05 , 问检验的结果如何?

解答: Bartlett 检验法

$$s^2 = \frac{\sum_{i=1}^k (n_i - 1) s_i^2}{N - k} = 101.25$$

$$C = 1 + \frac{1}{3(k - 1)} \left[\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{N - k} \right] = \frac{53}{48}$$

$$B = \frac{48}{53} \left[16 \ln_e 101.25 - \sum_{i=1}^k 4 \ln_e s_i^2 \right] = 2.945$$

$B = 2.945 < \chi^2_{0.05, 3} = 7.81$, 所以接受 H_0 。

12.6 参数估计

单因素方差分析, 参数的点估计及区间估计如表 12-6 所示。

表 12-6 单因素方差分析, 参数的点估计及区间估计

参数	符号	点估计	标准误差
方差	σ^2	MS_E	
总平均数	μ	\bar{y}	$\sqrt{\frac{MS_E}{N}}$
第 i 个处理平均数	$\mu_i = \mu + \alpha_i$	\bar{y}_i	$\sqrt{\frac{MS_E}{n_i}}$
第 i 个处理效应	α_i	$\bar{y}_i - \bar{y}$	$\sqrt{MS_E \left(\frac{N - n_i}{n_i N} \right)}$
处理平均数差	$\mu_i - \mu_j$	$\bar{y}_i - \bar{y}_j$	$\sqrt{\frac{MS_E}{n_i} + \frac{MS_E}{n_j}}$

参数的 $1 - \alpha$ 置信区间为: $[\text{点估计} \pm t_{\alpha/2}(N - a) (\text{标准误差})]$ 。

例题 12.6 (见网络资源)

12.7 双因素方差分析，无交互作用

随机化区组设计是用双因素方差分析，每格样本量为 1 的模型检验。在 Excel 或中文统计的菜单为“双因素方差分析：无重复试验”。

双因素方差分析，A × B 设计，每格样本量为 1，观测数据如表 12-7 所示。

表 12-7 观测数据

		因素 B				
		1	2	...	b	
因素 A	1	y_{11}	y_{12}	...	y_{1b}	$T_{1.}$
	2	y_{21}	y_{22}	...	y_{2b}	$T_{2.}$
	\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
	a	y_{a1}	y_{a2}	...	y_{ab}	$T_{a.}$
	Σ	$T_{.1}$	$T_{.2}$...	$T_{.b}$	$T_{..}$

数学模型如下：

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad i = 1, \cdots, a; \quad j = 1, \cdots, b$$

上述 i 是因素 A 中的第 i 种处理（共有 a 个处理），j 是因素 B 中的第 j 种处理（共有 b 个处理）。随机化区组设计是有一个因素是区组因素。

假定条件：

- (1) ε_{ij} 为互相独立，正态分布，平均数为 0，且有相同的方差， $\varepsilon_{ij} \sim N(0, \sigma^2)$ 。
 - (2) α_i 是因素 A 的第 i 处理的效应，是未知参数， $\sum_{i=1}^a \alpha_i = 0$ 。
 - (3) β_j 是因素 B 的第 j 处理的效应，是未知参数， $\sum_{j=1}^b \beta_j = 0$ 。
 - (4) Y_{ij} 为互相独立的正态分布，平均数为 $\mu + \alpha_i + \beta_j$ ，且有相同的方差 σ^2 （未知参数）， $Y_{ij} \sim N(\mu + \alpha_i + \beta_j, \sigma^2)$ 。
 - (5) 检验 $H_0^1: \alpha_1 = \alpha_2 = \cdots = \alpha_a, H_1^1: H_0^1$ 不成立，因素 A 效应显著。
检验 $H_0^2: \beta_1 = \beta_2 = \cdots = \beta_b, H_1^2: H_0^2$ 不成立，因素 B 效应显著。
- ANOVA 利用 F 检验，方差分析表如表 12-8 所示。

表 12-8 方差分析表（双因素方差分析，无交互作用）

变异来源	平方和	自由度	平均平方和	F 比值
因素 A	$SS_A = b \sum_{i=1}^a (\bar{y}_{i.} - \bar{\bar{y}})^2$	$a - 1$	$MS_A = \frac{SS_A}{a - 1}$	$F_1 = \frac{MS_A}{MS_E}$
因素 B	$SS_B = a \sum_{j=1}^b (\bar{y}_{.j} - \bar{\bar{y}})^2$	$b - 1$	$MS_B = \frac{SS_B}{b - 1}$	$F_2 = \frac{MS_B}{MS_E}$
误差	$SS_E = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{\bar{y}})^2$	$(a - 1) \times (b - 1)$	$MS_E = \frac{SS_E}{(a - 1)(b - 1)}$	
总和	$SS_T = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{\bar{y}})^2$	$ab - 1$		

若 $F_1 \geq F_{\alpha}[a - 1, (a - 1)(b - 1)]$ ，则拒绝 H_0^1 。若 $F_2 \geq F_{\alpha}[(b - 1), (a - 1)(b - 1)]$ ，则拒绝 H_0^2 。

双因素方差分析，无重复试验，参数的点估计及区间估计如表 12-9 所示。

表 12-9 参数的点估计及区间估计

参数	点估计	标准误差
方差 σ^2	MS_E	
总平均数	$\bar{\bar{y}}$	$\sqrt{MS_E/ab}$
因素 A 第 i 处理平均数	$\bar{y}_{i.}$	$\sqrt{MS_E/b}$
因素 B 第 j 处理平均数	$\bar{y}_{.j}$	$\sqrt{MS_E/a}$

参数的 $1 - \alpha$ 置信区间为： $\{\text{点估计} \pm t_{\alpha/2}[(a - 1)(b - 1)] \times (\text{标准误差})\}$

例题 12.7 表 12-10 是 10 个参加减肥计划者，6 周前与 6 周后的体重（单位千克）。

表 12-10 参加减肥计划者的体重变化

6 周前	70	103	78	64	73	96	84	55	74	90
6 周后	69	94	75	67	71	89	82	55	68	86

检验原假设：减肥计划在 6 周内对体重没有影响，即

$$H_0: \mu_1 - \mu_2 = 0$$

如果检验的显著性水平是 0.05，问检验的结果如何？

解答：如果用单因子方差分析，则方差分析表如表 12-11 所示。

表 12-11 方差分析表（单因子）

变异来源	自由度	平方和	平均平方和	F 比值
处理之间	1	$SS_A = 48.05$	$MS_A = 48.05$	$F = 0.267$
误差	18	$SS_E = 3242.5$	$MS_E = 180.14$	
总和 (Total)	19	$SS_T = 3290.55$		

因为 $F = 0.267 < F_{0.05}(1,18) = 4.414$ ，所以接受 H_0 ，处理（减肥计划）效应是不显著的。

注意： F 值是双总体“独立样本”检验的 t^* 值的平方。（比较例题 11.5）

如果用随机集区设计双因子方差分析，则方差分析表如表 12-12 所示。

表 12-12 方差分析表（双因子）

变异来源	自由度	平方和	平均平方和	F 比值
处理因素	1	$SS_A = 48.05$	$MS_A = 48.05$	$F_1 = 7.66$
集区因素	9	$SS_B = 3186.05$	$MS_B = 354.01$	$F_2 = 56.44$
误差	9	$SS_E = 56.45$	$MS_E = 6.27$	
总和 Total	19	$SS_T = 3290.55$		

因为 $F_1 = 7.66 > F_{0.05}(1,9) = 5.117$ ，所以否定 H_0^1 。处理（减肥计划）效应是显著的。

因为 $F_2 = 56.44 > F_{0.05}(9,9) = 3.179$ ，所以否定 H_0^2 。集区（个人）效应是显著的。

注意： F_1 值是双总体“相依样本”检验的 t^* 值的平方。（比较例题 11.6）

方差分析是将总变异分解为：因素 A 可解释的变异 $SSA = SS_A$ 和误差（因素 A 不可解释）的变异 $SSE(A)$ ，当前者比较大（如何比较），则因素 A 的影响效应是显著的。如果再加入因素 B，则将原来误差的变异 $SSE(A)$ 再分解为： $SS(B) = SS_B$ 和 $SSE(A, B)$ ，这样检验因素 A 和 B 的影响效应是显著的可能性就提高了，如图 12-5 所示。



图 12-5 方差分析增加因素，分解变异来源

方差分析是回归分析的一个特例，所以上述观念可以扩展到回归分析，如图 12-6 所示。

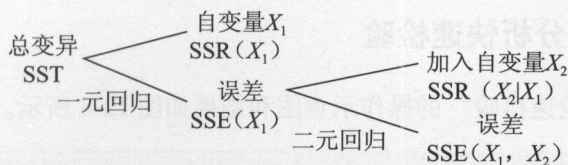


图 12-6 回归分析增加自变量，分解变异来源（请见第 13 章补充教材多元回归）

12.8 中文统计应用

12.8.1 单因素方差分析（例题 12.3）

执行“单因素方差分析”的操作示意图和结果如图 12-7 所示。



图 12-7 执行“单因素方差分析”的操作示意图和结果

12.8.2 方差分析快速检验

执行“方差分析快速检验”的操作示意图和结果如图 12-8 所示。

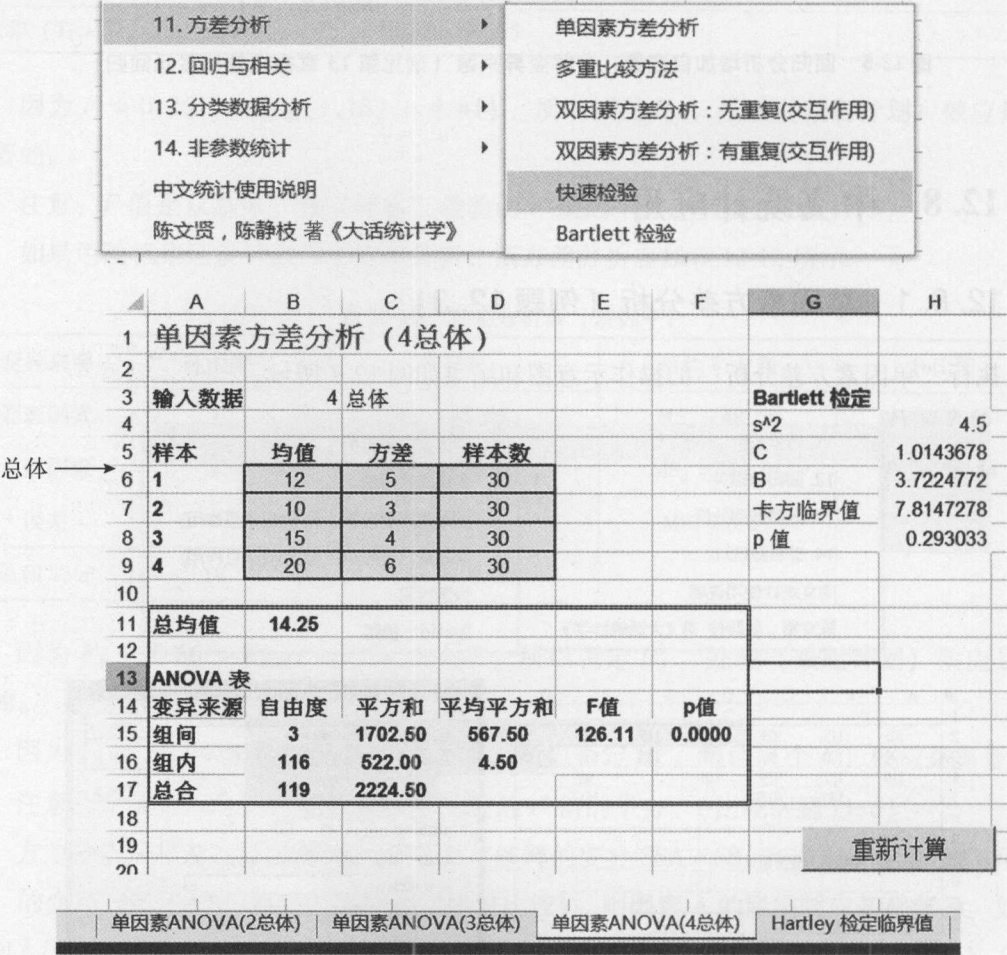


图 12-8 执行“方差分析快速检验”的操作示意图和结果

12.8.3 双因素方差分析：无重复（例题 12.9）

执行“双因素方差分析：无重复”的操作示意图和结果如图 12-9 所示。

11. 方差分析	▶	单因素方差分析
12. 回归与相关	▶	多重比较方法
13. 分类数据分析	▶	双因素方差分析：无重复(交互作用)
14. 非参数统计	▶	双因素方差分析：有重复(交互作用)
中文统计使用说明		快速检验
陈文贤, 陈静枝 著《大话统计学》		Bartlett 检验

	A	B	C	D
1				
2		4.7	9.4	6.3
3		3.5	7.6	5.1
4		0.1	5.3	1.8
5		1.6	6.2	3.6
6				
7				
8				
9				
10				
11				

双因素方差分析：无重复(交互作用)

输入

输入区域：

☐ 标志

显著性水平(a)：

输出选项

☐ 输出区域：

☒ 新工作表：

☐ 新工作

确定 取消 帮助

	A	B	C	D	E
1	方差分析：无重复双因素分析				
2					
3	UMMAR	观测数	求和	平均	方差
4	行 1	3	20.4	6.8	5.71
5	行 2	3	16.2	5.4	4.27
6	行 3	3	7.2	2.4	7.03
7	行 4	3	11.4	3.8	5.32
8					
9	列 1	4	9.9	2.475	4.135833
10	列 2	4	28.5	7.125	3.195833
11	列 3	4	16.8	4.2	3.78

14	方差分析					
15	差异源	SS	df	MS	F	P-value
16	行	32.88	3	10.96	144.5275	5.53E-06
17	列	44.205	2	22.1025	291.4615	1.06E-06
18	误差	0.455	6	0.075833		
19						
20	总计	77.54	11			

图 12-9 执行“双因素方差分析：无重复”的操作示意图和结果

12.9 本章流程图

本章流程图如图 12-10 所示。

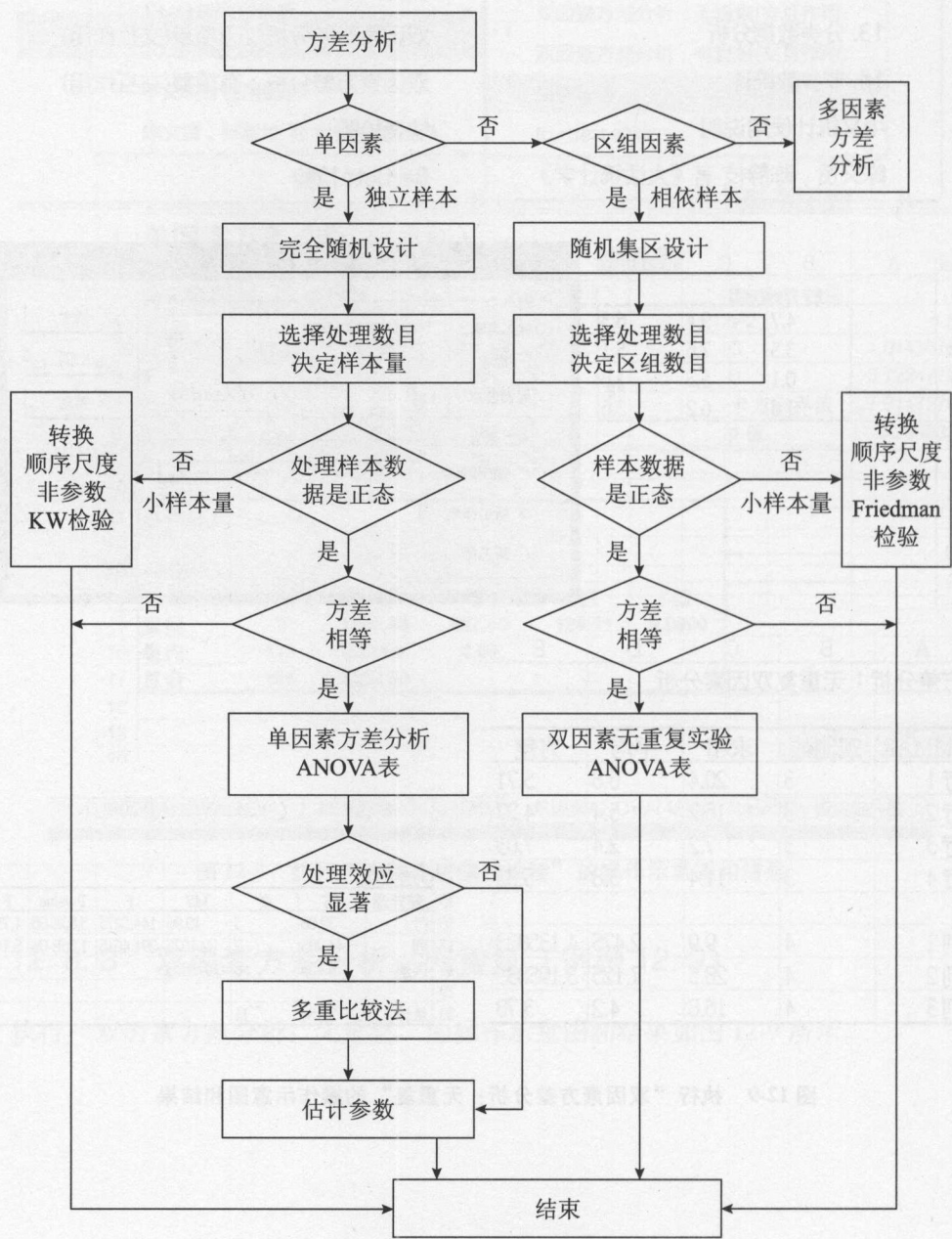


图 12-10 第 12 章流程图

12.11 本章思维导图

本章思维导图如图 12-11 所示。

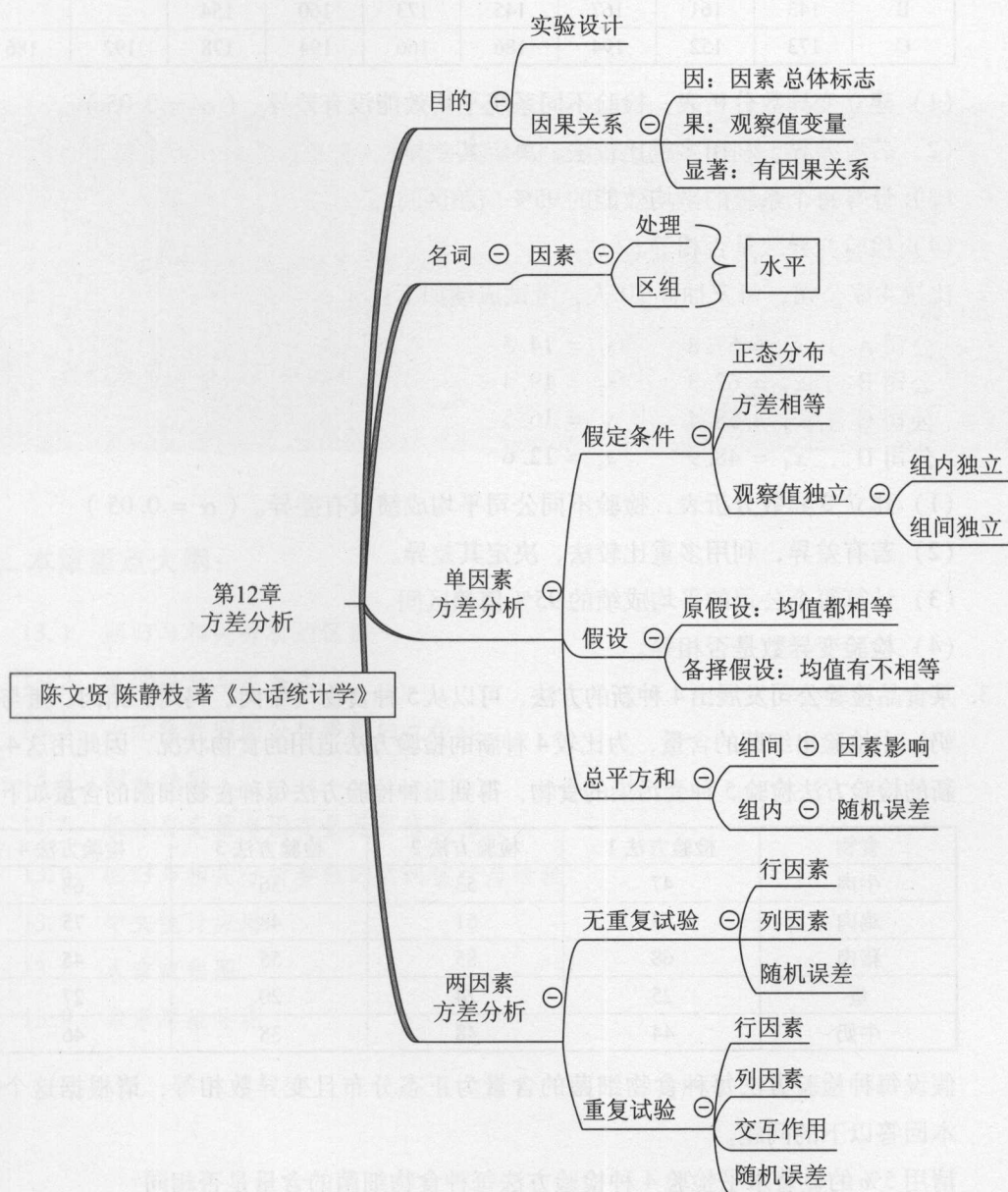


图 12-11 第 12 章思维导图

习题

1. 比较 3 种设计系统，效能如下：

A	147	188	162	144	157	179	165	180	
B	143	161	167	145	173	160	154		
C	173	152	194	186	166	194	178	192	186

- (1) 建立变异数分析表，检验不同系统平均效能没有差异。($\alpha = 0.05$)
 - (2) 若有差异，利用多重比较法，决定其差异。
 - (3) 计算每个系统的平均效能的 95% 信赖区间。
 - (4) 检验变异数是否相等。
2. 比较 4 家公司，每家抽样 30 人，考试成绩如下：

公司 A	$\bar{x}_1 = 57.8$	$s_1 = 14.3$
公司 B	$\bar{x}_2 = 62.3$	$s_2 = 19.4$
公司 C	$\bar{x}_3 = 58.4$	$s_3 = 16.5$
公司 D	$\bar{x}_4 = 48.9$	$s_4 = 12.6$

- (1) 建立变异数分析表，检验不同公司平均成绩没有差异。($\alpha = 0.05$)
 - (2) 若有差异，利用多重比较法，决定其差异。
 - (3) 计算每个公司的平均成绩的 95% 信赖区间。
 - (4) 检验变异数是否相等。
3. 某食品检验公司发展出 4 种新的方法，可以从 5 种食物（牛肉、鸡肉、猪肉、蛋与牛奶）中检验出细菌的含量，为比较 4 种新的检验方法适用的食物状况，因此用这 4 种新的检验方法检验 5 种受污染的食物，得到每种检验方法每种食物细菌的含量如下：

食物	检验方法 1	检验方法 2	检验方法 3	检验方法 4
牛肉	47	53	36	68
鸡肉	53	61	48	75
猪肉	68	85	55	45
蛋	25	24	20	27
牛奶	44	48	38	46

假设每种检验方法每种食物细菌的含量为正态分布且变异数相等，请根据这个样本回答以下的问题。

请用 5% 的显著水平检验 4 种检验方法每种食物细菌的含量是否相同？

其他习题请下载。



第 13 章

回归与相关分析

今据因果同时。若小乘说因果者，即转因以成果，因灭始果成。


——《华严一乘十玄门探玄》

有关系（检验结果显著），就没关系；没关系（不显著），就要再找关系。

——顺口溜

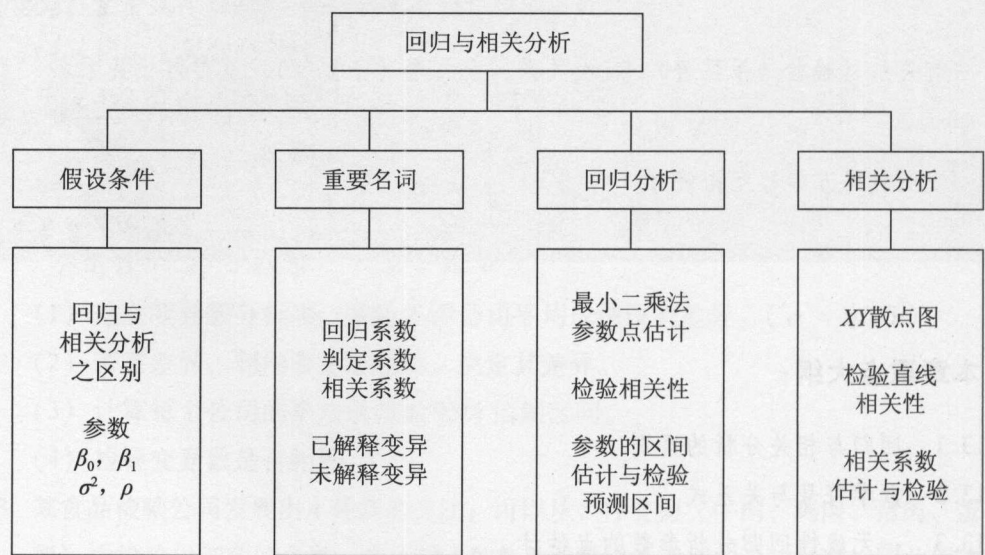
我们晓得万事都互相效力。

——《圣经·罗马书》



本章重点大纲：

- 13.1 回归与相关分析的区别
- 13.2 数学符号与关系式
- 13.3 一元线性回归分析参数的点估计
- 13.4 相关分析
- 13.5 检验自变量与因变量是否线性相关
- 13.6 回归与相关分析参数的区间估计与检验
- 13.7 中文统计应用
- 13.8 本章流程图
- 13.9 本章思维导图



本章概念图

13.1 回归与相关分析的区别

一元回归 (simple regression) 与相关分析 (correlation analysis) 都是探讨两个以上变量之间的“直线”关系。一元回归分析的主要目的,是探讨两个变量之间是否有直线的关系,并利用这个关系,来做预测。一元回归又称简单回归。

一元回归分析有两个变量,变量 X 与变量 Y ,通常我们将变量 X 称作自变量或独立变量 (independent variable),将变量 Y 称作因变量或依赖变量 (dependent variable)。 X 是因, Y 是果,一元回归分析要事先预设因果关系。所谓“一元”是只有一个自变量。“多元”是有两个以上自变量。

回归这名词出自 Sir Francis Galton 爵士 (1822—1911),他是著名的遗传学者,他研究父亲和儿子的身高,发现:父亲身高较平均数高的下一代比上一代矮,父亲身高较平均数矮的下一代比上一代高,他称这个现象为“回归平凡” (regression towards mediocrity)。以回归分析的术语来说:父亲身高是自变量 x ,儿子身高是因变量 y ,回归直线的斜率 (参数 β_1) 小于 1。(如果回归直线的斜率大于 1,则是高者恒高,矮者恒矮。)如图 13-1 所示。

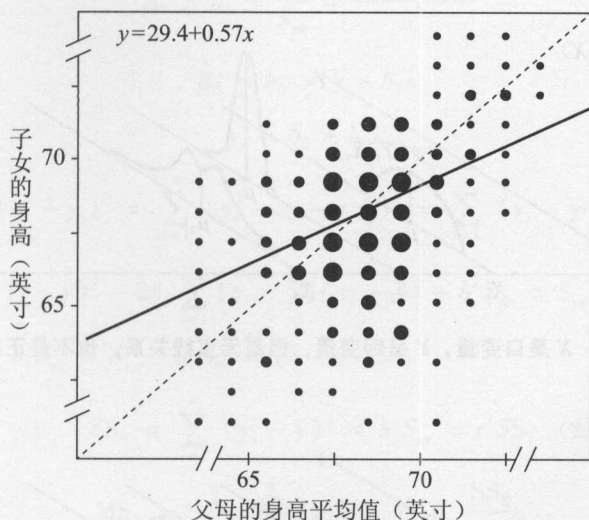


图 13-1 父亲与儿子的身高的关系

湖北省体育科学研究所,进行了研究校正:

儿子身高 = $(56.699 + 0.419 \times \text{父身高} + 0.265 \times \text{母身高}) \pm 3\text{cm}$

女儿身高 = $(40.089 + 0.306 \times \text{父身高} + 0.431 \times \text{母身高}) \pm 3\text{cm}$

相关分析，是检验两个变量之间是否有“线性相关性”以及相关的程度。相关性并“不显示因果关系”，从相关分析，我们并不知道两个变量，是 X 变量影响 Y 变量，或者 Y 变量影响 X 变量；也有可能第 3 个变量影响这两个变量，而造成它们之间的相关。

实际上，回归分析也不能“证明”或检验“因果”关系。如表 13-1 所示。

表 13-1 回归分析与相关分析的比较

回归分析	相关分析
相同点：检验两变量“线性”关系	相同点：检验两变量“线性”关系
不同点：	不同点：
1. X 是控制变量（自变量）， Y 是随机变量（因变量）， X, Y 先预设因果关系	1. X 是随机变量， Y 是随机变量 X, Y 没预设也无检验因果关系
2. 对每个 X_i 已知， Y 是正态分布，实际应记作 Y_{X_i} ，简记作 $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$	2. 对每个 X_i 已知， Y 是正态分布 对每个 Y_i 已知， X 是正态分布
3. 每个 Y_i 的正态分布的期望值是在一直线	3. X, Y 的联合分布是双变量正态分布
4. Y_i 有相同的方差 σ^2	4. X_i 有相同的方差 σ_X^2 Y_i 有相同的方差 σ_Y^2
5. S_{XX} 视为常数 S_{XY}, S_{YY} 是随机变量	5. X, Y 有相关系数 ρ S_{XX}, S_{XY}, S_{YY} 是随机变量
6. 主要参数 $\beta_0, \beta_1, \sigma^2$	6. 主要参数 ρ

图 13-2 不符合回归分析的假定条件。回归分析的 XY 变量关系图，如图 13-3 所示。

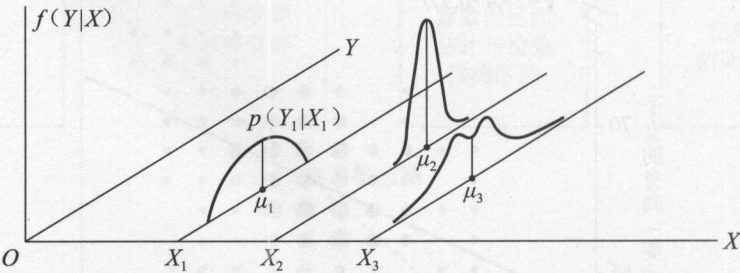


图 13-2 X 是自变量， Y 是因变量，但是无直线关系，也不是正态分布

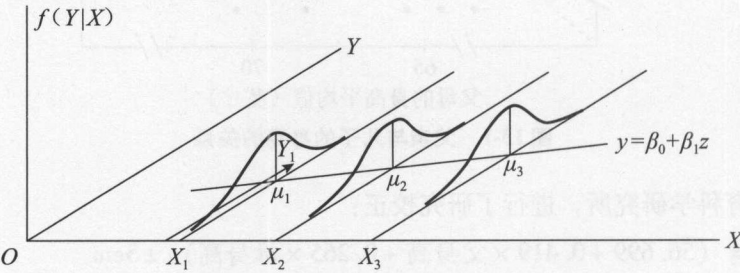


图 13-3 一元直线回归的 X, Y 变量关系图

13.2 数学符号与关系式

在本章,我们将用到一些平方和的符号与关系式,为了方便查阅,我们将它们集中在这节。请注意,以下 x 或 x_i 是常数值,而 y 或 y_i 代表因变量值。如果改为大写 Y 或 Y_i 则代表随机变量。例如 S_{xy} 为一个估计值,是一个实数值; S_{XY} 成为一个估计量,是一个随机变量。

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \dots, n$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$$

$$\text{Cov}(X, Y) = \frac{S_{xy}}{n-1}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2 \quad S_x^2 = \frac{S_{xx}}{n-1}$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 \quad S_y^2 = \frac{S_{yy}}{n-1}$$

$$\hat{\beta}_1 = b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\text{Cov}(X, Y)}{S_x^2}$$

$$\hat{\beta}_0 = b_0 = \bar{y} - b_1 \bar{x}$$

$$\hat{y}_i = b_0 + b_1 x_i$$

$$\begin{aligned} \text{SS}_E &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^n (y_i - \bar{y} - b_1 \bar{x} - b_1 x_i)^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2b_1 \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) + b_1^2 S_{xx} = S_{yy} - 2b_1 S_{xy} + b_1^2 S_{xx} \\ &= S_{yy} - b_1 S_{xy} = S_{yy} - b_1^2 S_{xx} = (1 - r^2) \text{SS}_T \end{aligned}$$

$$\text{SS}_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b_1^2 S_{xx} = r^2 \text{SS}_T$$

$$\text{MS}_E = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{\text{SS}_E}{n-2}$$

$$\text{SS}_T = \sum_{i=1}^n (y_i - \bar{y})^2 = \text{SS}_E + \text{SS}_R = S_{yy}$$

$$\hat{\sigma} = \sqrt{\frac{\text{SS}_E}{n-2}}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = b_1 \sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum (x_i - \bar{x})^2)(\sum (y_i - \bar{y})^2)}}$$

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

13.3 一元线性回归分析参数的点估计

一元线性回归分析的数学模型: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \dots, n$

假定条件:

(1) β_0, β_1 为未知参数。

(2) x_i 是人为选择变量, 控制变量, 自变量, 常数, 没有误差。

(3) ε_i 是误差项, 随机变量, 独立, 期望值为 0, 方差未知但相同。

$$E(\varepsilon_i) = 0, \quad V(\varepsilon_i) = \sigma^2, \quad \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$$

(4) Y_i 是随机变量, 独立, 期望值为 $\beta_0 + \beta_1 x_i$, 方差未知但相同。

$$E(Y_i) = \beta_0 + \beta_1 x_i, \quad V(Y_i) = \sigma^2, \quad \text{Cov}(Y_i, Y_j) = 0 \quad i \neq j$$

我们注意到以上假定并未提到 ε_i 和 Y_i 是正态分布, 因为我们目前只要对 β_0 和 β_1 做点估计, 等到做区间估计和检验时就需要正态分布的假定。

β_0 和 β_1 的点估计, 是利用最小二乘法 (least squares method)。这是使误差的平方和最小的估计值。令 b_0 和 b_1 分别是 β_0 和 β_1 的点估计。 b_0 和 b_1 是 β_0 和 β_1 的不偏估计量, $E(b_0) = \beta_0$, $E(b_1) = \beta_1$, 如图 13-4 所示。误差项的平方和是

$$SS_E = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

要使误差项的平方和最小, SS_E 分别对 b_0 和 b_1 做微分, 即

$$\frac{\partial SS_E}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i)$$

$$\frac{\partial SS_E}{\partial b_1} = -2 \sum_{i=1}^n x_i (y_i - b_0 - b_1 x_i)$$

令上述两个式子为 0, 得到下列正态 (或正规) 方程 (Normal equations)

$$\sum_{i=1}^n y_i = nb_0 + b_1 \sum_{i=1}^n x_i$$

$$\sum_{i=1}^n x_i y_i = b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2$$

解上述联立方程, 得

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = b_0 = \bar{y} - b_1 \bar{x}$$

因为 $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, 令 MS_E 为 σ^2 的点估计。 MS_E 是 σ^2 的不偏点估计量, 即

$$\hat{\sigma}^2 = MS_E = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{SS_E}{n-2}$$

$\sqrt{MS_E} = \sqrt{\hat{\sigma}^2} = \hat{\sigma}$ 是 σ 的点估计值, 称作回归标准误差。

直线回归的残差如图 13-4 所示。

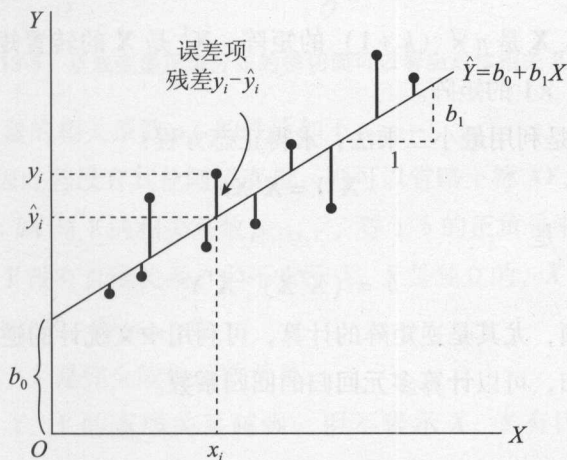


图 13-4 直线回归的残差

请注意: 点估计值 b_0, σ 的单位和 Y_i 的单位相同; b_1 的单位是 Y_i 的单位除以 X_i 的单位; 相关系数 r 是没有单位 (无名数) 的。例如: Y_i 的单位从千克改为克 (数据乘以 1000), X_i 的单位从米改为厘米 (数据乘以 100), 则回归直线截距 b_0 和 σ 改为乘以 1000, 回归直线斜率 b_1 改为乘以 10 (1000/100); 相关系数 r 没有改变, 检验直线相关的 p 值也不变。

点估计值 b_0, b_1 的公式来自解正态方程, 多元回归的正态方程如下:

(1) 多元回归方程 $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$, 去除 ε_i , Y_i 改为 y_i , β_j 改为 b_j 。

(2) 两边加 $\sum \Rightarrow \sum y_i = \sum b_0 + \sum b_1 x_{1i} + \sum b_2 x_{2i} + \cdots + \sum b_k x_{ki}$

(3) 上式两边分别乘以 $x_{1i}, x_{2i}, \cdots, x_{ki}$, 得到正态方程如下:

$$\sum y_i = nb_0 + b_1 \sum x_{1i} + b_2 \sum x_{2i} + \cdots + b_k \sum x_{ki}$$

$$\sum x_{1i}y_i = b_0 \sum x_{1i} + b_1 \sum x_{1i}^2 + b_2 \sum x_{1i}x_{2i} + \cdots + b_k \sum x_{1i}x_{ki}$$

$$\sum x_{2i}y_i = b_0 \sum x_{2i} + b_1 \sum x_{1i}x_{2i} + b_2 \sum x_{2i}^2 + \cdots + b_k \sum x_{2i}x_{ki}$$

$$\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots$$

$$\sum x_{ki}y_i = b_0 \sum x_{ki} + b_1 \sum x_{1i}x_{ki} + b_2 \sum x_{2i}x_{ki} + \cdots + b_k \sum x_{ki}^2$$

(4) 上述 $k+1$ 联立方程, 可以, 解出 $k+1$ 未知数 b_0, b_1, \cdots, b_k 。

用矩阵来表示:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \quad \mathbf{X}^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{pmatrix} \quad b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}$$

y 是 $n \times 1$ 的矩阵, \mathbf{X} 是 $n \times (k+1)$ 的矩阵, \mathbf{X}^T 是 \mathbf{X} 的转置矩阵, 是 $(k+1) \times n$ 的矩阵, b 是 $(k+1) \times 1$ 的矩阵。

β 的点估计 b , 也是利用最小二乘法, 求得正态方程:

$$\mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{X} b$$

所以 β 的估计量 b 是

$$b = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

矩阵的计算很麻烦, 尤其是逆矩阵的计算, 可利用中文统计的逆矩阵功能。中文统计的一元与多元线性回归, 可以计算多元回归的回归系数。

由于:

最小二乘法 \rightarrow SS_E 最小 \rightarrow 正态方程 \rightarrow 回归直线方程 \rightarrow $SS_T = SS_R + SS_E$

所以, $SS_T = SS_R + SS_E$ 并非都成立, 只有在最小二乘法的回归模式才成立。

当然, 方差分析 (ANOVA) 也成立, 因为方差分析是回归模式的一个特例。

例题 13.1 (见网络资源)

13.4 相关分析

相关分析, 是描述两个变量之间的“线性相关性”的程度。假定 X, Y 是两个随机变量, 定义总体相关系数 (population correlation coefficient) ρ

$$\rho = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

从双变量正态分布的横切面可以看出总体相关系数，如图 13-5 所示。

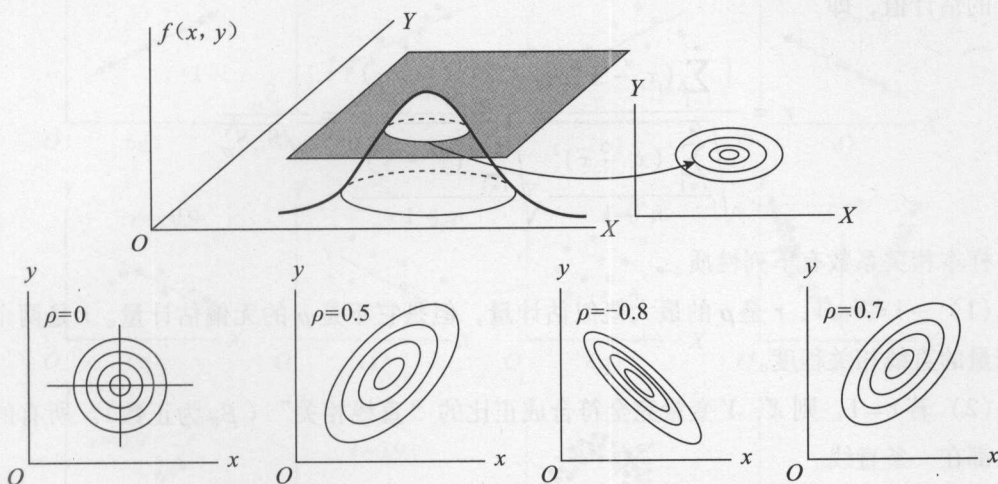


图 13-5 从双变量正态分布的横切面可以看出总体相关系数

X, Y 两个随机变量的相关系数 ρ_{XY} 的性质如下。

- (1) $\rho_{XY} = \rho_{YX}$ ，因此若没有其他随机变量，则可以省略下标 XY , $\rho_{XY} = \rho$ 。
- (2) 随机变量 $a + bX$ 与 Y 的相关系数 $\rho_{a+bX, Y}$ ，等于 b 的正负号乘以 ρ_{XY} 。
- (3) $\rho_{XY} = 0 \Rightarrow X, Y$ 没有直线关系，但不表示 X, Y 是独立的， X, Y 可能有曲线关系。
- (4) 若 X, Y 是独立的，则 $\rho_{XY} = 0$ 。
- (5) $\rho_{XY} = \pm 1 \Leftrightarrow X, Y$ 是完全线性函数关系。
- (6) $\rho_{XY} \rightarrow 1$ ，则 X, Y 的直线关系越强，但不表示 X, Y 有因果关系（例如回归关系）。

相关系数的关系如图 13-6 所示。

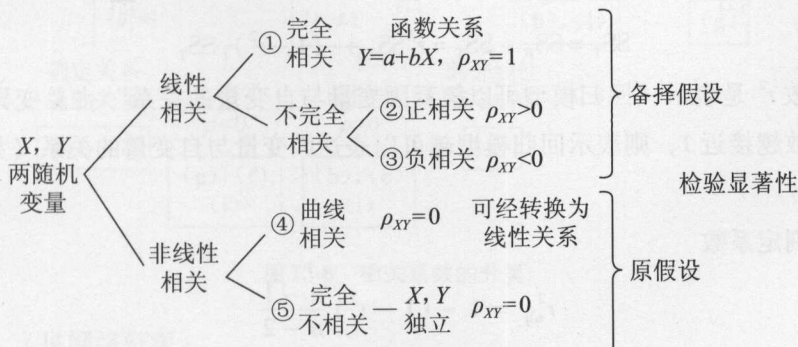


图 13-6 相关系数的分类

通常总体随机变量的相关系数是很难计算的， ρ 是未知参数，所以用样本数据计算。

X, Y 两个随机变量的样本相关系数 (sample correlation coefficient) r , 是总体相关系数 ρ 的估计值, 即

$$r = \frac{[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})](\frac{1}{n-1})}{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

样本相关系数有下列性质。

(1) $-1 \leq r \leq 1$, r 是 ρ 的最大似估计量, 但是它不是 ρ 的无偏估计量。 r 是两个随机变量的直线相关程度。

(2) 若 $r=1$, 则 X, Y 变量完全符合成正比的“直线相关” (β_1 为正数)。所有的样本点都在一条直线上。

(3) 若 $r=0$, 则 X, Y 变量完全无“直线相关” ($\beta_1 = 0$)。(可能有二次相关)

(4) 若 $r=-1$, 则 X, Y 变量完全符合成反比的“直线相关” (β_1 为负数)。

(5) r 的正负号表示 X, Y 的正相关与负相关。

(6) $|r|$ (r 的绝对值) 越大, 直线关系越强。

(7) r 的大小和回归直线的斜率无关, 只有其正负号和回归直线斜率的正负号相同。

图 13-7 (b), 图 13-7 (i), 图 13-7 (j) 是相同的样本相关系数, 不同的回归直线斜率。

判定系数 r^2 (coefficient of determination) 是 r 的平方

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

$$SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad SS_T = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$SS_T = SS_R + SS_E = r^2 SS_T + (1 - r^2) SS_T$$

判定系数 r^2 是表示“回归模型可以解释因变量与自变量的关系”占总变异的百分比。所以判定系数越接近 1, 则表示回归模型越可以表达因变量与自变量的关系 (如图 13-7 和图 13-8 所示)。

调整的判定系数

$$r_{\text{adj}}^2 = 1 - (1 - r^2) \frac{n-1}{n-2}$$

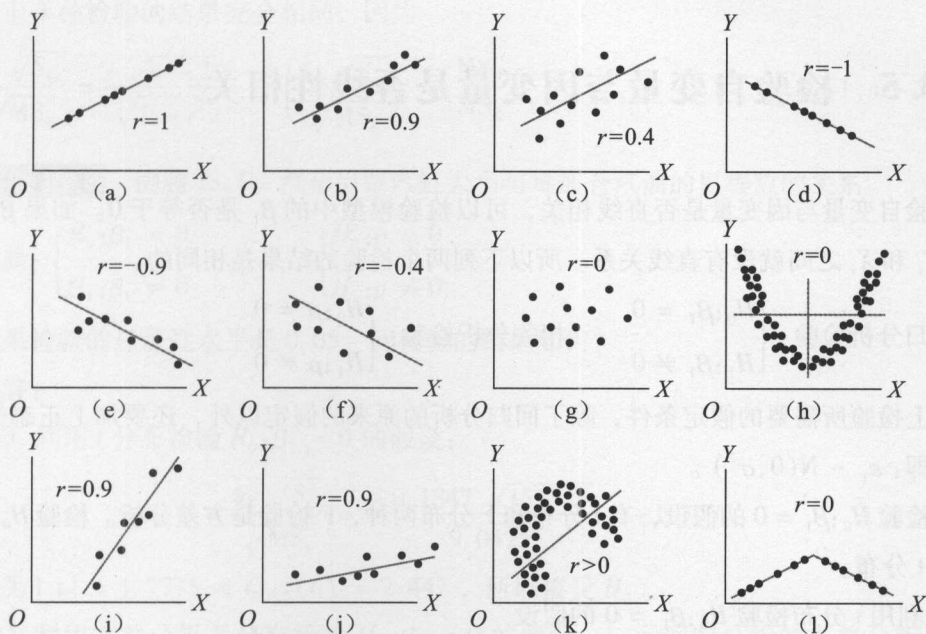


图 13-7 样本相关系数

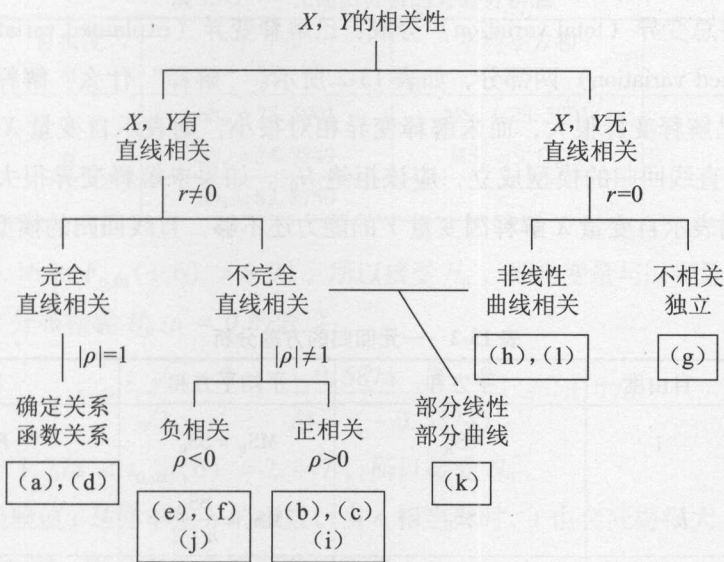


图 13-8 相关系数的分类

例题 13.2 (见网络资源)

13.5 检验自变量与因变量是否线性相关

检验自变量与因变量是否直线相关，可以检验模型中的 β_1 是否等于 0。如果 β_1 等于 0，则 Y_i 和 X_i 之间就没有直线关系。所以下列两个检验的结果是相同的。

回归分析检验：
$$\begin{cases} H_0:\beta_1 = 0 \\ H_1:\beta_1 \neq 0 \end{cases}$$

相关分析检验：
$$\begin{cases} H_0:\rho = 0 \\ H_1:\rho \neq 0 \end{cases}$$

以上检验所需要的假定条件，除了回归分析的原来的假定以外，还要加上正态分布的假定，即： $\varepsilon_i \sim N(0, \sigma^2)$ 。

要检验 $H_0:\beta_1 = 0$ 的假设，有 t 分布和 F 分布两种，F 检验是方差分析。检验 $H_0:\rho = 0$ 是利用 t 分布。

1. 利用 t 分布检验 $H_0:\beta_1 = 0$ 的假设

计算 $t = b_1 \sqrt{S_{xx}} / \sqrt{MS_E}$ ，若 $|t| \geq t_{\alpha/2}(n-2)$ ，则拒绝 H_0 。

2. 利用方差分析 F 分布检验 $H_0:\beta_1 = 0$ 的假设

回归分析将总变异（total variation）分成：已解释变异（explained variation）与未解释变异（unexplained variation）两部分，如表 13-2 所示。“解释”什么？解释回归模型是否“显著”。如果已解释变异很大，而未解释变异相对很小，则表示自变量 X 有足够的解释因变量 Y ，直线回归的模型成立，应该拒绝 H_0 。如果未解释变异很大，而已解释变异相对很小，则表示自变量 X 解释因变量 Y 的能力还不够，直线回归的模型不成立，应该接受 H_0 。

表 13-2 一元回归的方差分析

变异来源	自由度	平方和	平均平方和	F 比值
回归模型 (已解释变异)	1	SS_R	$MS_R = SS_R$	$F = \frac{MS_R}{MS_E}$
误差 (未解释变异)	$n - 2$	SS_E	$MS_E = \frac{SS_E}{n - 2}$	
总和	$n - 1$	SS_T		

若 $F \geq F_{\alpha}(1, n - 2)$ ，则拒绝 H_0 ，即自变量与因变量有显著关系。

3. 利用 t 分布检验 $H_0:\rho = 0$ 的假设

计算 $t = \frac{r \sqrt{n - 2}}{\sqrt{1 - r^2}}$ ， r 是样本相关系数。若 $|t| \geq t_{\alpha/2}(n - 2)$ ，则拒绝 H_0 。

以上3种检验的结果完全相同, 因为

$$\frac{b_1 \sqrt{S_{xx}}}{\sqrt{MS_E}} = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} \quad \left(\frac{b_1 \sqrt{S_{xx}}}{\sqrt{MS_E}} \right)^2 = \frac{MS_R}{MS_E} \quad [t_{\alpha/2}(n-2)]^2 = F_{\alpha}(1, n-2)$$

例题 13.3 例题 13.1, 汽车引擎汽缸大小与每加仑汽油的里程数的关系。

$$\text{检验: } \begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases} \quad \begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$$

如果检验的显著性水平是 0.05, 问检验的结果如何?

解答:

(1) 利用 t 分布检验 $H_0: \beta_1 = 0$ 的假设:

$$t = \frac{b_1 \sqrt{S_{xx}}}{\sqrt{MS_E}} = \frac{-0.1347 \sqrt{1575.5}}{\sqrt{9.0475}} = -1.7775$$

因为 $|t| = 1.7775 < t_{0.025}(6) = 2.447$, 所以接受 H_0 。

(2) 利用方差分析 F 分布检验 $H_0: \beta_1 = 0$ 的假设: 一元回归分析的方差分析表如表 13-3 所示。

表 13-3 一元回归分析的方差分析表

变异来源 Source	自由度 df	平方和 SS	平均平方和 MS	F 比值 F - ratio
回归模型	1	$SS_R = 28.5901$	$MS_R = 28.5901$	$F = 3.16$
误差	6	$SS_E = 54.2849$	$MS_E = 9.0475$	
总和	7	$SS_T = 82.8750$		

因为 $F = 3.16 < F_{0.05}(1, 6) = 5.99$, 所以接受 H_0 , 即自变量与因变量没有显著关系。

(3) 利用 t 分布检验 $H_0: \rho = 0$ 的假设:

$$t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.5874 \sqrt{8-2}}{\sqrt{1-(-0.5874)^2}} = -1.778$$

因为 $|t| = 1.778 < t_{0.025}(6) = 2.447$, 所以接受 H_0 。

请注意: 检验值 t 是样本量 n 的函数, 当 n 相当大时, t 也会变得很大。所以只要 $r \neq 0$, 当样本量很大时, 都会有显著的直线相关性。

13.6 回归与相关分析参数的区间估计与检验

一元线性回归分析的数学模型:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \cdots, n, \quad \varepsilon \sim N(0, \sigma^2)$$

从本节以后检验所需要的假设条件，除了回归分析的原来的 4 个假设以外，还要加上正态分布的假设，即： $\varepsilon_i \sim N(0, \sigma^2)$ 。

一元线性回归与相关分析的有关分布如表 13-4 所示。

表 13-4 一元线性回归与相关分析的有关分布

参数	估计量	有关分布
β_0	b_0	$\sqrt{\frac{nS_{xx}}{MS_E \sum_{i=1}^n x_i^2}} (b_0 - \beta_0) \sim t(n-2)$
β_1	b_1	$\sqrt{\frac{S_{xx}}{MS_E}} (b_1 - \beta_1) \sim t(n-2)$
$\beta_0 + \beta_1 x_p$	$b_0 + b_1 x_p$	$\frac{b_0 + b_1 x_p - \beta_0 - \beta_1 x_p}{\sqrt{MS_E \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right]}} \sim t(n-2)$
y_p	$b_0 + b_1 x_p$	$\frac{y_p - b_0 - b_1 x_p}{\sqrt{MS_E \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right]}} \sim t(n-2)$
σ^2	$\frac{SS_E}{n-2}$	$\frac{SS_E}{\sigma^2} \sim \chi^2(n-2)$
ρ	r	$\frac{r \sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2) \quad (\text{当 } \rho \text{ 接近 } 0)$

13.6.1 方差的区间估计

回归模型中的方差 σ^2 是未知参数，其估计很重要，因为后续的估计和检验，都要用到方差 σ^2 的估计值。

回归模型中 Y_i 与 ε_i 的共同方差 σ^2 的点估计（不偏估计量）是 MS_E

$$MS_E = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{SS_E}{n-2}$$

方差 σ^2 的 $1 - \alpha$ 置信区间

$$\frac{SS_E}{\chi^2_{\alpha/2}(n-2)} \leq \sigma^2 \leq \frac{SS_E}{\chi^2_{1-\alpha/2}(n-2)}$$

例题 13.4 （见网络资源）

13.6.2 回归参数 β_0 的区间估计与检验

回归参数 β_0 的估计量 b_0 ，是一个随机变量，根据以上假设， b_0 是一个正态分布， $b_0 \sim$

$N(\beta_0, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n(\sum_{i=1}^n x_i^2 - n\bar{x}^2)})$ 。在回归分析， β_0, σ^2 为未知参数； $\sum x_i^2, S_{xx}$ 为已知常数。

$$(1) \text{ 令 } S_{b_0}^2 = \frac{SS_E}{n-2} \left[\frac{\sum_{i=1}^n x_i^2}{nS_{xx}} \right] = MS_E \left(\frac{\sum_{i=1}^n x_i^2}{nS_{xx}} \right)$$

由于 $\sqrt{\frac{nS_{xx}}{MS_E \sum_{i=1}^n x_i^2}} (b_0 - \beta_0) \sim t(n-2)$ ，所以

$$S_{b_0} = \sqrt{MS_E \left(\frac{\sum_{i=1}^n x_i^2}{nS_{xx}} \right)}$$

(2) β_0 的区间估计， β_0 的 $1-\alpha$ 置信区间

$$b_0 - t_{\alpha/2}(n-2)S_{b_0} \leq \beta_0 \leq b_0 + t_{\alpha/2}(n-2)S_{b_0}$$

(3) β_0 的检验

$$\text{双侧: } \begin{cases} H_0^I: \beta_0 = c \\ H_1^I: \beta_0 \neq c \end{cases} \quad \text{左侧: } \begin{cases} H_0^{II}: \beta_0 \geq c \\ H_1^{II}: \beta_0 < c \end{cases} \quad \text{右侧: } \begin{cases} H_0^{III}: \beta_0 \leq c \\ H_1^{III}: \beta_0 > c \end{cases}$$

(4) 计算检验值

$$t = \frac{b_0 - c}{S_{b_0}}$$

若 $|t| \geq t_{\alpha/2}(n-2)$ ，则拒绝 H_0^I ；若 $t < -t_{\alpha}(n-2)$ ，则拒绝 H_0^{II} ；若 $t > t_{\alpha}(n-2)$ ，则拒绝 H_0^{III} 。

13.6.3 回归参数 β_1 的区间估计与检验

回归参数 β_1 的估计量 b_1 ，是一个随机变量，根据上述假设， b_1 是一个正态分布， $b_1 \sim$

$$N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2})$$

$$(1) \text{ 令 } S_{b_1}^2 = \frac{SS_E}{(n-2)(\sum_{i=1}^n x_i^2 - n\bar{x}^2)} = \frac{MS_E}{S_{xx}}$$

由于 $\sqrt{\frac{S_{xx}}{MS_E}}(b_1 - \beta_1) \sim t(n-2)$, 所以

$$S_{b_1} = \sqrt{\frac{MS_E}{S_{xx}}} = \sqrt{\frac{MS_E}{\sum_i x_i^2 - n\bar{x}^2}}$$

(2) β_1 的区间估计, β_1 的 $1 - \alpha$ 置信区间:

$$b_1 - t_{\alpha/2}(n-2)S_{b_1} \leq \beta_1 \leq b_1 + t_{\alpha/2}(n-2)S_{b_1}$$

(3) β_1 的检验:

$$\text{双侧: } \begin{cases} H_0^I: \beta_1 = c \\ H_1^I: \beta_1 \neq c \end{cases} \quad \text{左侧: } \begin{cases} H_0^{II}: \beta_1 \geq c \\ H_1^{II}: \beta_1 < c \end{cases} \quad \text{右侧: } \begin{cases} H_0^{III}: \beta_1 \leq c \\ H_1^{III}: \beta_1 > c \end{cases}$$

(4) 计算检验值

$$t = \frac{b_1 - c}{S_{b_1}}$$

若 $|t| \geq t_{\alpha/2}(n-2)$, 则拒绝 H_0^I ; 若 $t < -t_{\alpha}(n-2)$, 则拒绝 H_0^{II} ; 若 $t > t_{\alpha}(n-2)$, 则拒绝 H_0^{III} 。

13.6.4 回归系数 β_1 的估计量 b_1 的讨论

b_1 的方差的讨论:

$$V(b_1) = \frac{\sigma^2}{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right)} = \frac{\sigma^2}{(n-1)\left[\sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)\right]} = \frac{\sigma^2}{(n-1)S_x^2}$$

式中: S_x^2 ——自变量 x_i 的方差, 即

$$S_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)}$$

要使 β_1 的估计量 b_1 越准确, 就要减小 $V(b_1)$, 如图 13-9 所示。要减小 $V(b_1)$, 有下列 3 种方法。

- (1) 减小 σ , 即减小因变量 Y_i 的变异程度。
- (2) 增加样本量 n 。
- (3) 增加 S_x^2 , 即自变量 x_i 的方差。因为 x 是控制变量, 所以它的方差是可以控制, 只要扩大 x 的范围。

例题 13.5 (见网络资源)

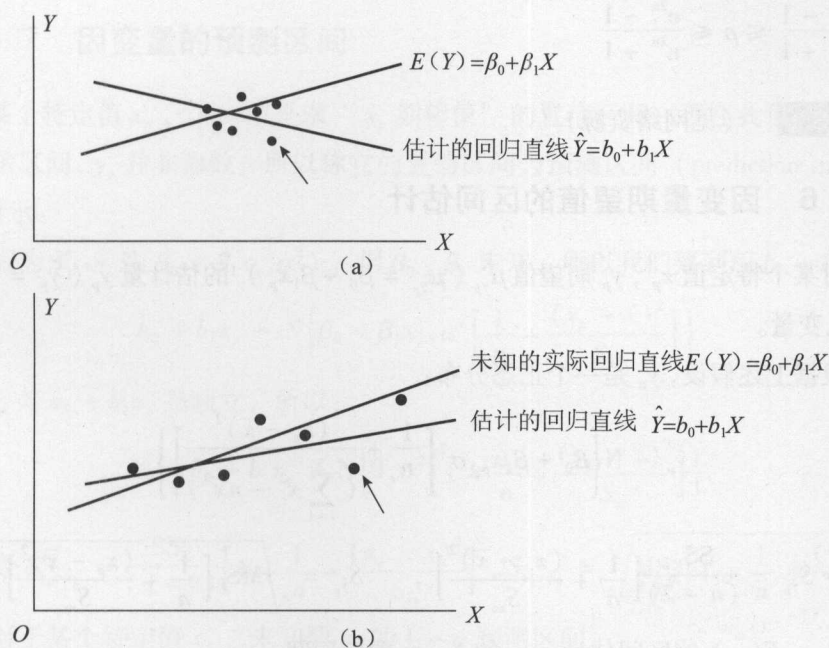


图 13-9 方差与 β_1 的估计量之间的关系

(a) 因为 x_i 的方差小, 所以 β_1 的估计不准确; (b) 因为 x_i 的方差大, 所以 β_1 的估计较准确

13.6.5 相关系数的区间估计与检验

两个随机变量的样本相关系数 r , 是总体相关系数 ρ 的估计量 (或估计值, 符号相同)。 r 是 ρ 的最大似然估计量, 但是它不是 ρ 的无偏估计量。

(1) 当 ρ 接近 $+1$, r 的分布是左偏分布。 ρ 的置信区间并非以 r 为中心左右对称, 而是 r 的右边 (大于的部分) 较小, r 的左边 (小于的部分) 较大。

(2) 当 ρ 接近 -1 , r 的分布是右偏分布。 ρ 的置信区间并非以 r 为中心左右对称, 而是 r 的右边 (大于的部分) 较大, r 的左边 (小于的部分) 较小。

(3) 当 ρ 等于 0 , r 的分布是 t 分布, 自由度为 $n-2$ 。 ρ 的置信区间, 是以 r 为中心左右对称。

ρ 的 $1-\alpha$ 置信区间计算, 要经过 Fisher 转换, 其计算步骤如下。

- (1) 计算样本相关系数 r 。
- (2) 计算 Fisher 转换

$$Z_r = \frac{1}{2} \ln_e \left(\frac{1+r}{1-r} \right)$$

$$(3) \quad l = Z_r - z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}} \quad u = Z_r + z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n-3}}$$

(4) $\frac{e^{2l}-1}{e^{2l}+1} \leq \rho \leq \frac{e^{2u}-1}{e^{2u}+1}$

例题 13.6 (见网络资源)

13.6.6 因变量期望值的区间估计

- (1) 对某个特定值 x_p, y_p 期望值 μ_{y_p} ($\mu_{y_p} = \beta_0 + \beta_1 x_p$) 的估计量 \hat{y}_p ($\hat{y}_p = b_0 + b_1 x_p$) 是一个随机变量。
- (2) 根据上述假设, \hat{y}_p 是一个正态分布:

$$\hat{y}_p \sim N\left\{\beta_0 + \beta_1 x_p, \sigma^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{(\sum_{i=1}^n x_i^2 - n\bar{x}^2)} \right] \right\}$$

(3) 令 $S_{\hat{y}_p}^2 = \frac{SS_E}{(n-2)} \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right], \quad S_{\hat{y}_p} = \sqrt{MS_E \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right]}$

- (4) $\mu_{y_p} = E(\hat{y}_p)$ 的区间估计, μ_{y_p} 的 $1 - \alpha$ 置信区间:

$$\hat{y}_p - t_{\frac{\alpha}{2}, n-2} S_{\hat{y}_p} \leq \mu_{y_p} \leq \hat{y}_p + t_{\frac{\alpha}{2}, n-2} S_{\hat{y}_p}$$

- (5) μ_{y_p} 的检验

$$\begin{cases} H_0^I: \mu_{y_p} = c \\ H_1^I: \mu_{y_p} \neq c \end{cases} \quad \begin{cases} H_0^{II}: \mu_{y_p} \geq c \\ H_1^{II}: \mu_{y_p} < c \end{cases} \quad \begin{cases} H_0^{III}: \mu_{y_p} \leq c \\ H_1^{III}: \mu_{y_p} > c \end{cases}$$

- (6) 计算检验值

$$t = \frac{\hat{y}_p - c}{S_{\hat{y}_p}}$$

若 $|t| \geq t_{\frac{\alpha}{2}, n-2}$, 则拒绝 H_0^I ; 若 $t < -t_{\frac{\alpha}{2}, n-2}$, 则拒绝 H_0^{II} ; 若 $t > t_{\frac{\alpha}{2}, n-2}$, 则拒绝 H_0^{III} 。如果 13-10 所示。

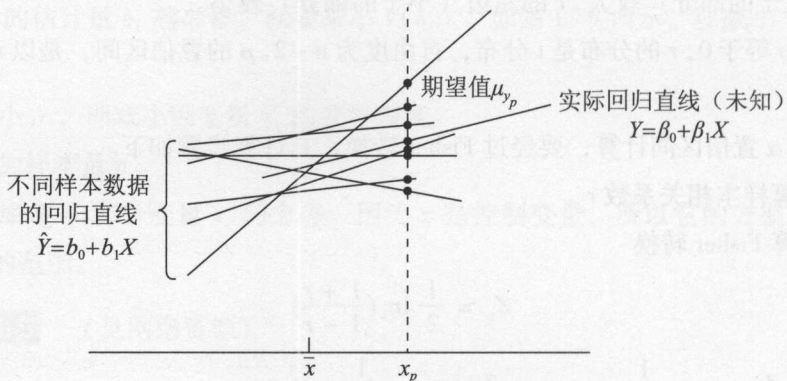


图 13-10 预测值 \hat{y}_p 的抽样分布

13.6.7 因变量的预测区间

对于某个特定值 x_p ，上一节是求“ y_p 期望值”的置信区间，现在我们要计算“实际 y_p ”的置信区间， y_p 并非参数，所以称它的置信区间为预测区间 (prediction interval)，如图 13-11 所示。

(1) 因为 $Y_p \sim N(\beta_0 + \beta_1 x_p, \sigma^2)$ ，但 β_0, β_1 未知，所以我们要利用 $b_0 + b_1 x_p$ ，而

$$b_0 + b_1 x_p \sim N\left\{\beta_0 + \beta_1 x_p, \sigma^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right]\right\}$$

(2) Y_p 与 $b_0 + b_1 x_p$ 是独立，所以：

$$Y_p - b_0 - b_1 x_p \sim N\left\{0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right]\right\}$$

(3) 令 $S_{y_p - \hat{y}_p}^2 = \frac{SS_E}{(n-2)} \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right]$ ， $S_{y_p - \hat{y}_p} = \sqrt{MS_E \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right]}$

(4) 对于某个特定值 x_p ，未知数 y_p 的 $1 - \alpha$ 预测区间：

$$b_0 + b_1 x_p - t_{\alpha/2}(n-2) S_{y_p - \hat{y}_p} \leq y_p \leq b_0 + b_1 x_p + t_{\alpha/2}(n-2) S_{y_p - \hat{y}_p}$$

(5) y_p 的检验

$$\text{双侧: } \begin{cases} H_0^I: y_p = c \\ H_1^I: y_p \neq c \end{cases} \quad \text{左侧: } \begin{cases} H_0^{II}: y_p \geq c \\ H_1^{II}: y_p < c \end{cases} \quad \text{右侧: } \begin{cases} H_0^{III}: y_p \leq c \\ H_1^{III}: y_p > c \end{cases}$$

(6) 计算检验值

$$t = \frac{b_0 + b_1 x_p - c}{S_{y_p - \hat{y}_p}}$$

若 $|t| \geq t_{\alpha/2}(n-2)$ ，则拒绝 H_0^I ；若 $t < -t_{\alpha}(n-2)$ ，则拒绝 H_0^{II} ；若 $t > t_{\alpha}(n-2)$ ，则拒绝 H_0^{III} 。

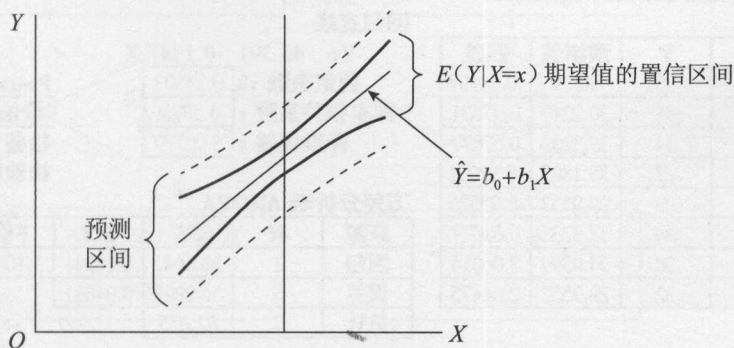


图 13-11 预测区间与期望值的置信区间

13.7 中文统计应用

一元线性回归（例题 13.1）

一元线性回归输入数据如图 13-12 所示。执行“一元线性回归”的操作示意图和结果如图 13-13 所示。

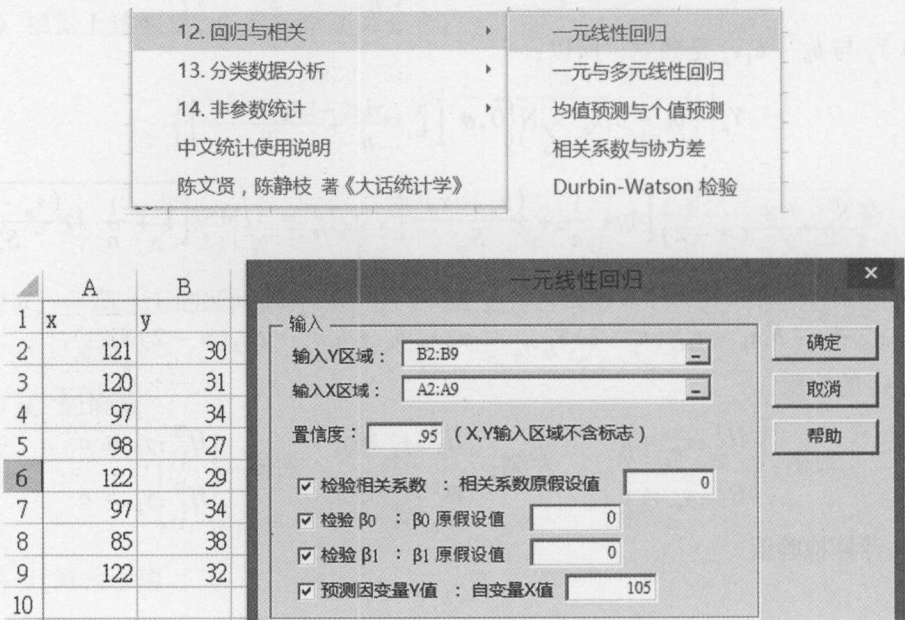


图 13-12 一元线性回归输入数据



图 13-13 执行“一元线性回归”的操作示意图和结果

13.8 本章流程图

本章流程图如图 13-14 所示。

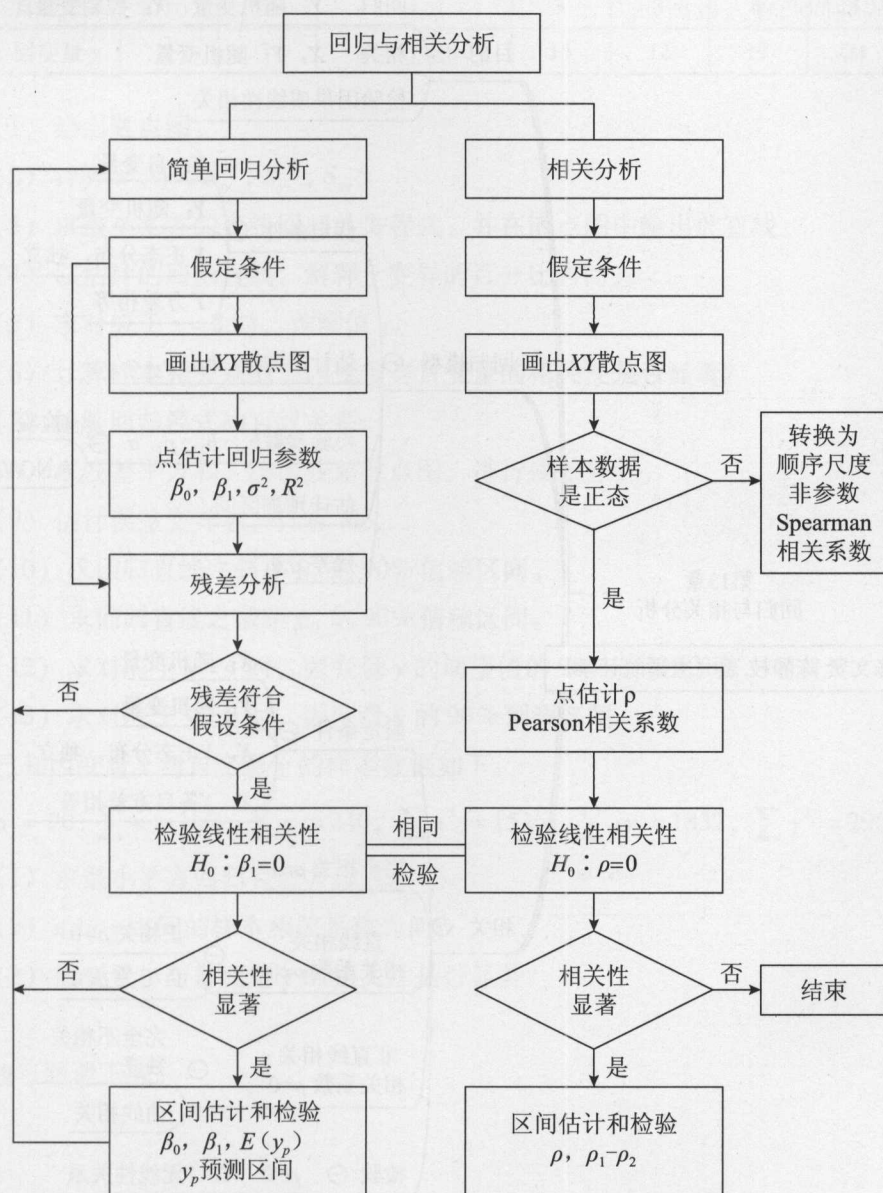


图 13-14 第 13 章流程图

13.9 本章思维导图

本章思维导图如图 13-15 所示。

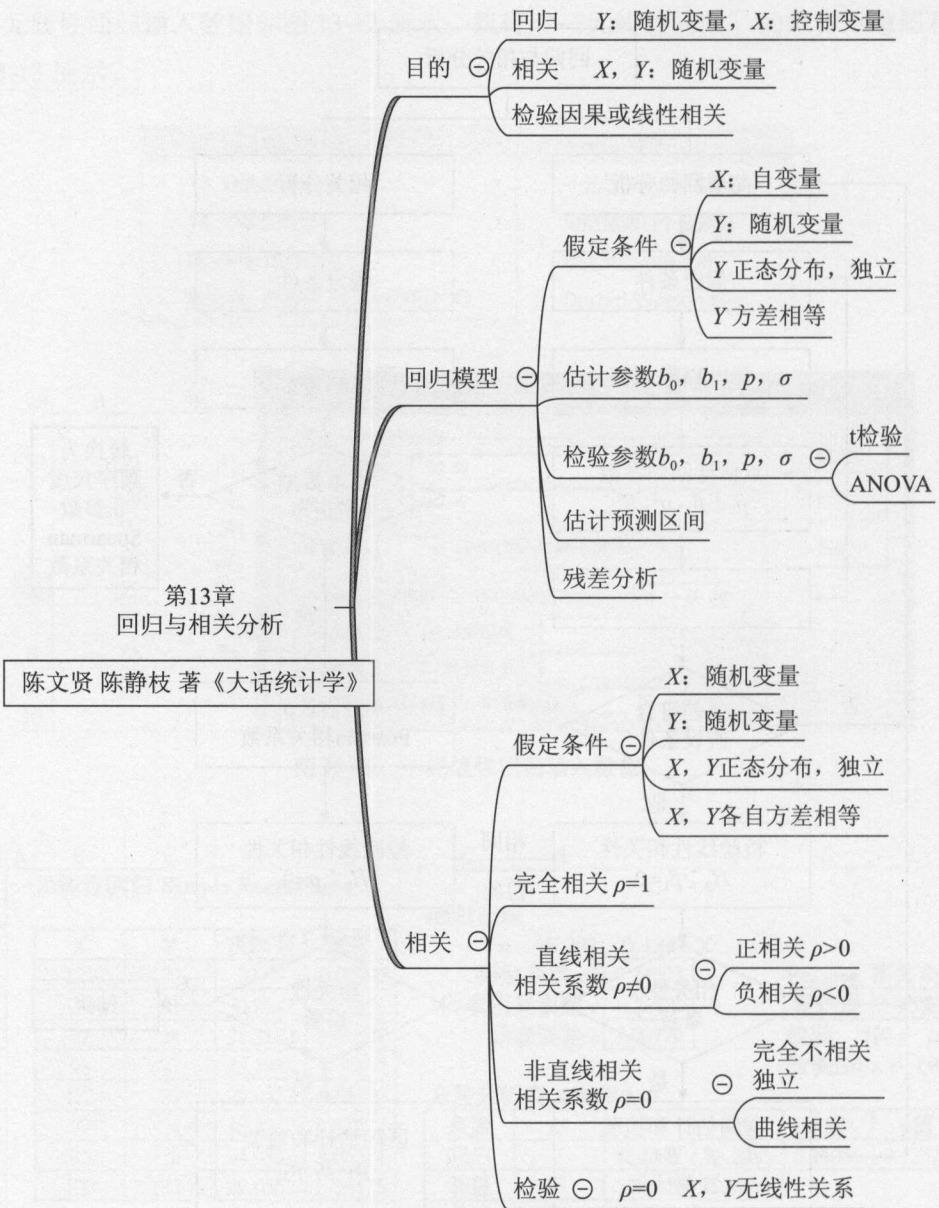


图 13-15 第 13 章思维导图

习 题

1. 已知下列9组 (x, y) 值:

自变量 x	1	1	1	2	3	3	4	5	5
因变量 y	9	7	8	10	15	12	19	24	21

- (1) 绘出散点图。
 - (2) 计算 \bar{x} , \bar{y} , S_{xx} , S_{yy} , S_{xy} 。
 - (3) 求最小平方的直线回归的方程式, 并在散点图中绘出此直线。
 - (4) 以估计的回归直线, 解释 y 变异的百分比为何?
 - (5) 求对应于 $x=3$ 时, 预测值 y 。
 - (6) 计算样本相关系数。因变量与自变量的相关性是否显著?
 - (7) 检验回归模式的直线关系。
 - (8) 求残差平方和, 绘出残差散点图, 进行残差分析。
 - (9) 估计误差变异数, 计算 MS_E 。
 - (10) 求回归直线之斜率 β_1 的 90% 信赖区间。
 - (11) 求回归直线之截距 β_0 的 90% 信赖区间。
 - (12) 求对应于 $x=4$ 时, 因变量 y 的期望值的 90% 信赖区间。
 - (13) 求对应于 $x=6$ 时, 因变量 y 的 90% 预测区间。
2. 已知因变量 x 与自变量 y 的样本数据如下:

$$n = 20, \sum x = 160, \sum y = 240, \sum x^2 = 1536, \sum xy = 1832, \sum y^2 = 2965$$

- (1) 求最小平方回归直线方程式。
- (2) x 与 y 之间的样本相关系数为何?
- (3) 因变量 x 与自变量 y 的相关性是否显著?

其他习题请下载。



第 14 章

分类数据分析

道生一，一生二，二生三，三生万物。

——老子《道德经》

方以类聚，物以群分，吉凶生矣。

——《易经·系辞上》

(引申为：物以类聚，人以群分)

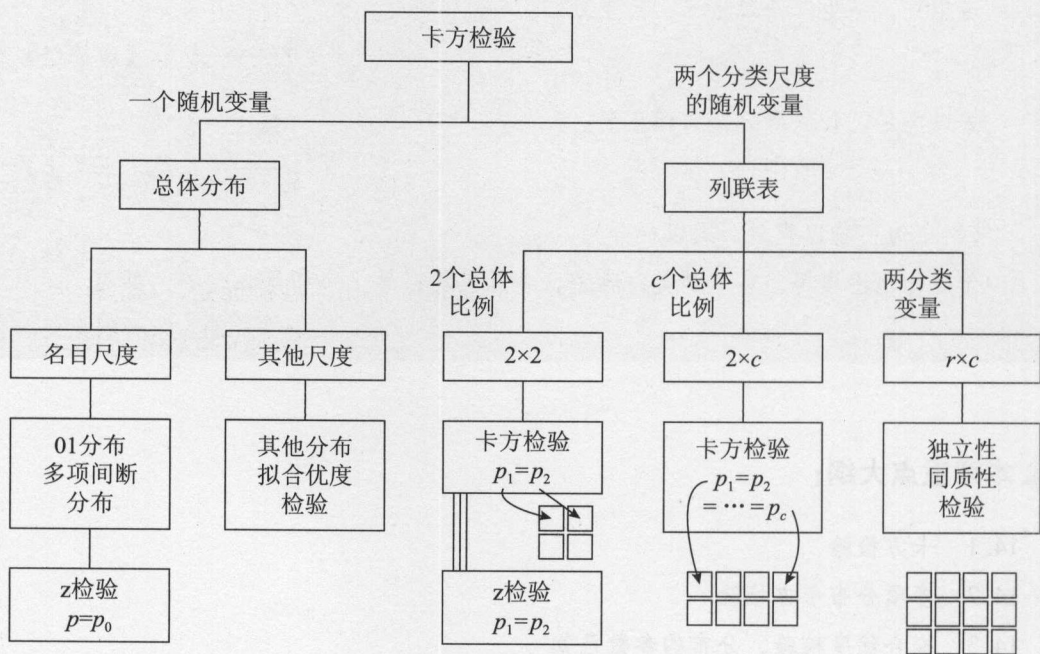
于是神造出野兽，各从其类；牲畜，各从其类；地上一切昆虫，各从其类。

——《圣经·创世纪》



本章重点大纲：

- 14.1 卡方检验
- 14.2 多项分布卡方检验
- 14.3 拟合优度检验，分布的参数已知
- 14.4 拟合优度检验，分布的参数未知
- 14.5 卡方检验独立性与同构性
- 14.6 中文统计应用
- 14.7 本章流程图
- 14.8 本章思维导图



本章概念图

14.1 卡方检验

分类数据分析是利用卡方分布 (chi-square distribution) 作为检验的根据, 称为卡方检验 (chi-square test)。在第 10 章 10.8 节单总体 (正态分布) 检验方差, 用到了卡方分布。

卡方分布定义于所有大于等于 0 的正实数, 其参数是自由度 n , 它是右偏型分布。

在本章的分类数据分析, 是以卡方分布, 应用在下列两种检验。

- (1) 一个总体随机变量的概率分布是否服从某个分布? (拟合优度检验)
- (2) 一个总体两个名目 (分类) 尺度的随机变量是否独立? (独立性检验)

卡方检验, 主要是计算: 样本值 (O) 减理论值 (E) 的平方除以理论值 (E), 加以总和, 即

$$\chi^2 = \sum \frac{(O - E)^2}{E} = \sum \frac{(\text{本值} - \text{理值})^2}{\text{理值}}$$

样本值和理论值都是正整数 (计算次数)。理论值: 根据原假设成立, 在总样本量目 (或列联表的边际总和) 之下, 理论上应该出现的次数。然后, 再和卡方临界值比较, 决定检验结果。其中要注意卡方分布的自由度。

14.2 多项分布卡方检验

多项分布卡方检验 (multinomial chi-square test) (如图 14-1 所示), 是检验样本数据是否服从某一特定的离散概率分布, 即 5.1.1 节的任意离散型概率分布。

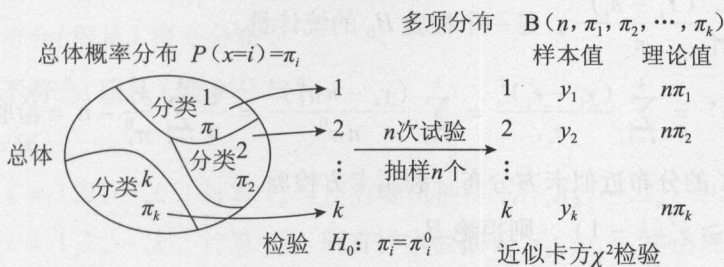


图 14-1 多项分布卡方检验

二项分布的检验 $\sum \frac{(\text{本值} - \text{理论值})^2}{\text{理论值}} = \frac{(x - n\pi)^2}{n\pi(1 - \pi)}$, 是总体比例检验值 z^* 的平方,

而 $\chi^2_{\alpha}(1) = (z_{\alpha/2})^2$ ，所以二项分布的总体比例检验（如图 14-2 所示）是多项分布卡方检验的特例。

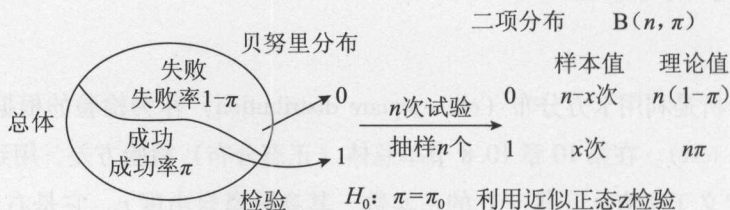


图 14-2 二项分布正态检验

1) 多项分布卡方检验的假设条件。

(1) 总体随机变量的值域为: $\{1, 2, \dots, k\}$ ，即随机变量只出现在: $1, 2, \dots, k$ ，其概率分别为: $\pi_1, \pi_2, \dots, \pi_k$ ，是未知参数。 $\sum \pi_i = 1$ 。

(2) X_1, X_2, \dots, X_n 为总体随机变量的随机抽样。

(3) 利用抽样的样本数据，检验：

$$\begin{cases} H_0: \pi_1 = \pi_1^0, \pi_2 = \pi_2^0, \dots, \pi_k = \pi_k^0 \\ H_1: \text{以上至少有一不等式} \end{cases}$$

2) 检验步骤。

(1) 对于 $i = 1, 2, \dots, k$ ， Y_i = 所有随机抽样 $\{X_1, X_2, \dots, X_n\}$ 中出现 i 的次数。

(2) 对于 $i = 1, 2, \dots, k$ ，计算 y_i = 所有样本数据 $\{x_1, x_2, \dots, x_n\}$ 中出现 i 的次数， y_i = 是抽样出现的次数。

(3) $\sum_{i=1}^k y_i = n$ = 样本的总数。令 $e_i = n\pi_i^0 = n$ 个样本出现 i 的期望值， e_i = 理论出现的次数（理论值）。 y_i = 抽样出现的次数（样本值）。

(4) $K = \sum_{i=1}^k \frac{(Y_i - e_i)^2}{e_i}$ ，是一个检验 H_0 的统计量。

(5) 计算 $k^* = \sum_{i=1}^k \frac{(y_i - e_i)^2}{e_i} = \sum_{i=1}^k \frac{(y_i - n\pi_i^0)^2}{n\pi_i^0} = \frac{1}{n} \sum_{i=1}^k \frac{y_i^2}{\pi_i^0} - n$ = 检验 H_0 的统计值。

(6) 以上 K 的分布近似卡方分布，利用卡方检验。

(7) 若 $k^* \geq \chi^2_{\alpha}(k-1)$ ，则拒绝 H_0 。

(8) p 值 = $P\{\chi^2(k-1) \geq k^*\}$ ，若 p 值 $\leq \alpha$ ，则拒绝 H_0 。

3) 注意事项。

(1) 当样本量 $n \leq 25$ ，以连续的卡方分布，来近似离散的样本数据，要做连续性修正：

$$k^* = \sum_{i=1}^k \frac{(|y_i - e_i| - 0.5)^2}{e_i}$$

(2) 有的统计学者建议, 每组 (即每个 i) 的期望值 (即 $e_i = n\pi_i^0$), 要大于 5。如果有小于 5 的组, 就要和它邻近的组合并。(k 的数目减 1, $k = k - 1$)

$Y_i = n$ 次抽样中出现 i 的次数。 Y_1, Y_2, \dots, Y_k 的联合离散概率分布称为 “多项分布” (multinomial distribution), 记作 $B(n, p_1, p_2, \dots, p_k)$, 其概率分布函数为

$$P(Y_1 = n_1, Y_2 = n_2, \dots, Y_k = n_k) = \frac{n!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}$$

$$\sum_{i=1}^k n_i = n$$

当 $k=2$ 即为二项分布。

14.3 拟合优度检验, 分布的参数已知

拟合优度检验 (chi-square goodness-of-fit test), 是检验样本数据是否服从, 某一特定的概率分布 F 。 F 可以是连续型概率分布, 但是要将其定义域分为 k 个区间。

F 是表示一般的分布, 而不是 F 分布 $F(m, n)$ 。

1) 拟合优度检验的假设条件。

(1) 总体随机变量为 X , 其值域分成: (A_1, A_2, \dots, A_k) 等互斥且周延的 k 组。如果 X 为离散型分布, 则 A_i 可能是整数; 如果 X 为连续型分布, 则 A_i 是实数区间。

(2) 将概率分布 F , 计算 $P(X \in A_i) = P(A_i)$ 为 A_i 出现的概率。

(3) X_1, X_2, \dots, X_n 为总体随机变量的随机抽样。

(4) 利用抽样的样本数据, 检验:

$$\begin{cases} H_0: \text{总体符合(服从) 概率分布 } F \\ H_1: \text{总体不符合(服从) 概率分布 } F \end{cases}$$

2) 检验步骤。

(1) 对于 $i = 1, 2, \dots, k$, 计算 $Y_i =$ 所有随机抽样 $\{X_1, X_2, \dots, X_n\}$ 中出现 A_i 的次数。

(2) 对于 $i = 1, 2, \dots, k$, 计算 $y_i =$ 所有样本数据 x_1, x_2, \dots, x_n 中出现 A_i 的次数, y_i 是抽样出现的次数。因此, 样本的总次数为

$$\sum_{i=1}^k y_i = n$$

(3) 令 $P(A_i)$ 为 A_i 出现的概率。 $e_i = nP(A_i) = n$ 个样本出现 A_i 的期望值。 e_i 是理论出

现的次数。

- (4) $K = \sum_{i=1}^k \frac{(Y_i - e_i)^2}{e_i}$ ，是一个检验 H_0 的统计量。
- (5) 计算 $k^* = \sum_{i=1}^k \frac{(y_i - e_i)^2}{e_i} = \sum_{i=1}^k \frac{[y_i - nP(A_i)]^2}{nP(A_i)}$ = 检验 H_0 的统计值。
- (6) 以上 K 的分布近似卡方分布，利用卡方检验。
- (7) 若 $k^* \geq \chi^2_{\alpha}(k-1)$ ，则拒绝 H_0 。
- (8) p 值 = $P\{\chi^2(k-1) \geq k^*\}$ ，若 p 值 $\leq \alpha$ ，则拒绝 H_0 。

例题 14.1 一名推销员每天访问 5 家，记录每天推销成功的家数。表 14-1 是 100 天的推销概况：每天推销成功的家数，服从二项分布 $B(5, 0.35)$ 上述不成立。

表 14-1 100 天的推销概况

推销成功的家数	0	1	2	3	4	5
出现天数	15	21	40	14	6	4

市场营销专家认为这名推销员每天推销成功的家数，应该服从二项分布 $B(5, 0.35)$ 。

检验： $\begin{cases} H_0: \text{每天推销成功的家数, 服从二项分布 } B(5, 0.35) \\ H_1: \text{上述不成立} \end{cases}$

如果检验的显著性水平是 0.05，问检验的结果如何？

解答：计算二项分布 $B(5, 0.35)$ 的概率如表 14-2 所示。

表 14-2 二项分布 $B(5, 0.35)$ 的概率

y	$P(y)$	$e = 100P(y)$
0	0.1160	11.60
1	0.3124	31.24
2	0.3364	33.64
3	0.1812	18.12
4	0.0487	4.87
5	0.0053	0.53

因为 $y=5$ 的期望值小于 1，所以要和 $y=4$ 合并为一组，如表 14-3 所示。

表 14-3 合并后, 二项分布 B (5, 0.35) 的概率

A_i	y_i	e_i	$y_i - e_i$	$(y_i - e_i)^2$	$(y_i - e_i)^2 / e_i$
0	15	11.60	3.40	11.5600	0.9966
1	21	31.24	-10.24	104.8576	3.3565
2	40	33.64	6.36	40.4496	1.2024
3	14	18.12	-4.12	16.9744	0.9368
4, 5	10	5.40	4.60	21.1600	3.9185
					10.4108

因为 $k^* = 10.4108 > \chi_{0.05,4}^2 = 9.488$, 所以拒绝 H_0 。 p 值 = 0.034。

14.4 拟合优度检验, 分布的参数未知

拟合优度检验, 是检验样本数据是否服从, 某一特定的概率分布 F, 但是其参数未知。

1) 拟合优度检验的假设条件。

(1) 总体随机变量为 X , 其值域分成: $\{A_1, A_2, \dots, A_k\}$ 等 k 组。但是总体的概率分布参数未知。

(2) 概率分布 F, $P(X \in A_i) = P(A_i)$ 为 A_i 出现的概率, 因为 F 的参数未知, 所以 $P(A_i)$ 暂时还未知。

(3) X_1, X_2, \dots, X_n 为总体随机变量的随机抽样。

(4) 利用抽样的样本数据, 检验: $\begin{cases} H_0: \text{总体符合(服从) 概率分布 F} \\ H_1: \text{总体不符合(服从) 概率分布 F} \end{cases}$

2) 检验步骤。

(1) 将样本数据, 对分布 F 的未知参数做点估计, 代入概率计算 $P(A_i)$ 。例如泊松分布, 用 λ 的点估计 \bar{x} 来计算概率 $P(A_i)$ 。

(2) 对于 $i = 1, 2, \dots, k$, Y_i = 所有样本数据 $\{X_1, X_2, \dots, X_n\}$ 中出现 A_i 的次数。

对于 $i = 1, 2, \dots, k$, 计算 y_i = 所有样本数据 $\{x_1, x_2, \dots, x_n\}$ 中出现 A_i 的次数, y_i 是抽样出现的次数。

(3) 计算 $e_i = nP(A_i)$, e_i 是理论出现的次数。

(4) $K = \sum_{i=1}^k \frac{(Y_i - e_i)^2}{e_i}$, 是一个检验 H_0 的统计量。

- (5) 计算 $k^* = \sum_{i=1}^k \frac{(y_i - e_i)^2}{e_i} = \sum_{i=1}^k \frac{[y_i - nP(A_i)]^2}{nP(A_i)}$ = 检验 H_0 的统计值。
- (6) 如果概率分布 F 有 m 个参数，因为以估计值代替参数，所以自由度要再减 m 。
- (7) 若 $k^* \geq \chi^2_\alpha(k - m - 1)$ ，则拒绝 H_0 。
- (8) p 值 = $P\{\chi^2(k - m - 1) \geq k^*\}$ ，若 p 值 $\leq \alpha$ ，则拒绝 H_0 。

例题 14.2 如果每页打字打错字的比率相同。以下是 100 页抽样中每页错字的数目：

每页错字的数目	0	1	2	3	4	5	6
出现页数	13	24	31	18	11	2	1

统计学者认为每页错字数目，应该服从泊松分布 $Pois(\lambda)$ ， λ 未知。

检验： $\begin{cases} H_0: \text{每页错字数目服从泊松分布 } Pois(\lambda) \\ H_1: \text{上述不成立} \end{cases}$

如果检验的显著性水平是 0.05，问检验的结果如何？

解答：上述 100 页抽样数据，每页平均错字是 2 个字，即

$$(24 + 2 \times 31 + 3 \times 18 + 4 \times 11 + 5 \times 2 + 6 \times 1) / 100 = 2$$

计算泊松分布 $Pois(2)$ 的概率如表 14-4 所示。

表 14-4 泊松分布 $Pois(2)$ 的概率

Y	$P(y)$	$e = 100P(y)$
0	0.1353	13.53
1	0.2707	27.07
2	0.2707	27.07
3	0.1804	18.04
4	0.0902	9.02
5	0.0361	3.61
6	0.0120	1.20
>6	0.0045	0.45

因为 $y=5, y=6, y>6$ 3 组的 e_i 都小于 5，所以将 5, 6 和大于 6 合并为一组，表 14-5 合并后，泊松分布 $Pois(2)$ 的概率。

表 14-5 合并后，泊松分布 Poiss (2) 的概率

A_i	y_i	$P(A_i)$	e_i
0	13	0.1353	13.53
1	24	0.2707	27.07
2	31	0.2707	27.07
3	18	0.1804	18.04
4	11	0.0902	9.02
>4	3	0.0527	5.26

因为 $k^* = 2.35 < \chi^2_{0.05}(4) = 9.488$ ，所以接受 H_0 。 p 值 = 0.67。

例题 14.3 统计学的 50 位学生的成绩如下：85，72，69，88，56，61，68，80，77，72，81，93，66，87，79，67，83，90，66，78，76，85，82，88，89，63，65，88，90，60，63，62，87，56，72，68，83，87，90，92，95，66，63，57，64，76，62，70，80，90。请检验是否服从正态分布？显著性水平 0.05。

解答：原假设正态分布的平均数和标准差未知，利用样本数据，得平均数 75.74，标准差 11.487。

将正态分布的实数定义域分为 4 组，如表 14-6 所示。

表 14-6 正态分布的实数定义域

标准常态区间	成绩分数区间	概率	期望次数 e_i	观察次数 y_i
$Z \leq -1$	$(-\infty, 64.25)$	0.1587	7.94	11
$-1 < Z \leq 0$	$(64.25, 75.74)$	0.3413	17.07	12
$0 < Z \leq 1$	$(75.74, 87.23)$	0.3413	17.07	16
$Z \geq 1$	$(87.23, +\infty)$	0.1587	7.94	11

$$k^* = \sum_{i=1}^k \frac{(y_i - e_i)^2}{e_i} = \frac{(11 - 7.94)^2}{7.94} + \frac{(12 - 17.07)^2}{17.07} + \frac{(16 - 17.07)^2}{17.07} + \frac{(11 - 7.94)^2}{7.94} = 3.94$$

因为 $k^* = 3.94 > \chi^2_{0.05}(1) = 3.8415$ ，所以拒绝 H_0 ，不是正态分布。 p 值 = 0.0471。

卡方分布自由度只有 1，因为 $k - m - 1 = 4 - 2 - 1 = 1$ ，两个参数用样本数据。

本例题可用“中文统计”：分类数据分析→正态分布检验。

14.5 卡方检验独立性与同构性

独立性卡方检验 (chi - square test of independent)，是检验样本数据，在二分类列联表 (two - way cross contingency table) 中，每个交集事件是否独立。若二分类代表两随机变量，则是检验两者是否独立。

同构性卡方检验 (chi - square test of homogeneity)，将二分类联立事件表中某一类，视作独立总体，检验这些总体对应另一类中有相同的性质 (比例)。例如两总体比例相同。

如果是独立性卡方检验，则将一个总体的 n 个随机抽样，分成 A, B 两大类, y_{ij} 为其每格的出现数目。检验两种分类是否独立的。

独立性检验和同构性检验，计算过程完全相同，检验的卡方分布，自由度也相同，其差别在表 14-7 说明。

表 14-7 独立性检验和同构性检验之比较

比较	独立性	同构性 (齐一性)
观测数据	列联表	列联表
列与行	一个总体抽样一组样本 (两个变量) 栏 (列): 分类变量 行: 分类变量	不同总体各抽样一组样本 (一个变量) 栏 (列): 多组总体 行: 分类变量
原假设	两个分类变量是独立的	每个总体的分类比例是相同的
实验设计 (调查的设计)	只能固定总样本量	可以固定列或行的边际和 (每个总体的样本量)
例如	教育程度和结婚次数 性别和收入 星座和主修科系 (或职业) 血型和职业	不同地区 (总体) 的政党 (分类) 支持度 不同教师 (总体) 的工作满意度

例如，教育程度分为：大学以上程度、高中以下程度；结婚次数分为：未婚、结婚一次、结婚两次以上；决定街头随机抽样 500 人 (总样本量)，调查以后填入列联表。性别当然是男女，收入分为几个范围，虽然可以固定抽样男性或女性的样本量，不过如果用电话随机访问，只要他 (她) 愿意回答，都列入样本量。

如果是同构性卡方检验，则 $R_i, i = 1, \cdots, a$ 为第 i 个总体的随机抽样值的数目。如图 14-3 所示，将 A 类视作不同总体, y_{ij} 为第 i 个总体出现 B 类 j 组的数目。检验各总体的 B 类比例是否相同。

二分类列联表，观察数据如图 14-3 所示。

		分类B				
		1	2	...	b	总和
分类A	1	y_{11}	y_{12}	...	y_{1b}	R_1
	2	y_{21}	y_{22}	...	y_{2b}	R_2
	⋮	⋮	⋮	⋮	⋮	⋮
	a	y_{a1}	y_{a2}	...	y_{ab}	R_a
	总和	C_1	C_2	...	C_b	n

图 14-3 观察数据

列联表独立性的条件:

$$y_{ij} \times n = R_i \times C_j \quad \text{或} \quad y_{ij} = \frac{R_i C_j}{n} \quad \forall i, j$$

独立性卡方检验，是利用抽样的样本数据，检验:

- $\begin{cases} H_0: \text{分类 A 与分类 B 是独立的} \\ H_1: \text{分类 A 与分类 B 不是独立的} \end{cases}$

同构性卡方检验，是利用抽样的样本数据，检验:

- $\begin{cases} H_0: \text{A 类每组对 B 类有一致性比例} \\ H_0: (y_{11}:y_{12}:\cdots:y_{1b}) = (y_{21}:y_{22}:\cdots:y_{2b}) = \cdots = (y_{a1}:y_{a2}:\cdots:y_{ab}) \quad (\text{每行有相同比例}) \\ (y_{11}:y_{21}:\cdots:y_{a1}) = (y_{12}:y_{22}:\cdots:y_{a2}) = \cdots = (y_{1b}:y_{2b}:\cdots:y_{ab}) \quad (\text{每列有相同比例}) \\ \text{或} \frac{y_{i1}}{C_1} = \frac{y_{i2}}{C_2} = \frac{y_{i3}}{C_3} = \cdots = \frac{y_{ib}}{C_b}, \frac{y_{1i}}{R_1} = \frac{y_{2i}}{R_2} = \frac{y_{3i}}{R_3} = \cdots = \frac{y_{ai}}{R_a} \quad \forall i \\ H_1: \text{以上至少有一不等式} \end{cases}$

例如: 如图 14-4 所示。

		分类 B			总和
		1	2	3	
分 类 A	1	2	3	5	10
	2	4	6	10	20
	3	6	9	15	30
	总和	12	18	30	60

图 14-4 举例

独立性: $2 \times 60 = 12 \times 10$, $3 \times 60 = 18 \times 10$, $5 \times 60 = 30 \times 10$, $4 \times 60 = 12 \times 20$, ...,

15 × 60 = 30 × 30

同构性：2 : 3 : 5 = 4 : 6 : 10 = 6 : 9 : 15, 2 : 4 : 6 = 3 : 6 : 9 = 5 : 10 : 15

2/12 = 3/18 = 5/30, 4/12 = 6/18 = 10/30, 6/12 = 9/18 = 15/30,

2/10 = 4/20 = 6/30, 3/10 = 6/20 = 9/30, 5/10 = 10/20 = 15/30

所以，上述列联表，分类 A 与分类 B 是独立的，有同构性。

独立性与同构性卡方检验的方法步骤都相同。

检验步骤：(R_i = 第 i 行和, C_j = 第 j 列和)

(1) 计算 $R_i = \sum_{j=1}^b y_{ij}, C_j = \sum_{i=1}^a y_{ij}, n = \sum_{i=1}^a \sum_{j=1}^b y_{ij}, e_{ij} = \frac{R_i C_j}{n}$ 。

(2) $K = \sum_{i=1}^a \sum_{j=1}^b \frac{(Y_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^a \sum_{j=1}^b \frac{Y_{ij}^2}{e_{ij}} - n$ ，是一个检验 H_0 的统计量。

(3) 计算 $k^* = \sum_{i=1}^a \sum_{j=1}^b \frac{(y_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^a \sum_{j=1}^b \frac{y_{ij}^2}{e_{ij}} - n$ = 检验 H_0 的统计值。

(4) 若 $k^* \geq \chi^2_{\alpha}[(a - 1) \times (b - 1)]$ ，则拒绝 H_0 。

如果 k^* 大到一定程度，表示样本数据和独立性的差异程度够大，显著非独立，有相关。

(5) 有关自由度的说明：一个三角形的三个角总和是 180° （固定的），所以两个角已知就决定第3个角，于是自由度是2。一个列联表，每列总和与每行总和固定，则 $(a - 1)(b - 1)$ 个格子的观察值已知，就可以决定其他所有格子的观察值。例如：一个 2×3 的二分类事件表，如果 y_{11} 与 y_{12} 已知，则其他值可以由每行每列之和算出，所以只有2个自由度。

例题 14.4 1912 年泰坦尼克号撞上冰山而沉没，乘客和组员共 2223 人，死亡 1517 人，其中不同“性别”（因）的死亡率（果）是否有显著差异；不同“身份（旅客等级或组员）”的“死亡率”，是否有显著差异？显著性水平是 0.05。（表 14-8 中的因果关系不是统计推论，是研究者判断）

表 14-8 泰坦尼克号生死录

因 \ 果	头等舱		二等舱		三等舱		组员（船员服务生）		总和	
	男	女	男	女	男	女	男	女	男	女
存活	54	145	15	104	69	105	194	20	332	374
	199		119		174		214		706	
死亡	119	11	142	24	417	119	682	3	1360	157
	130		166		536		685		1517	
总和	173	156	167	128	486	224	876	23	1692	531
	329		285		710		899		2223	

解答:

(1) 身份和存活率的独立性检验:

H_0 : 身份和存活率是独立的

$$k^* = \sum_{i=1}^a \sum_{j=1}^b \frac{y_{ij}^2}{e_{ij}} - n = \frac{2223 (199)^2}{(329)(706)} + \frac{2223 (119)^2}{(285)(706)} + \frac{2223 (174)^2}{(710)(706)} + \frac{2223 (214)^2}{(899)(706)} \\ + \frac{2223 (130)^2}{(329)(1517)} + \frac{2223 (166)^2}{(285)(1517)} + \frac{2223 (536)^2}{(710)(1517)} + \frac{2223 (685)^2}{(899)(1517)} - 2223 \\ = 181.89$$

$k^* = 181.89 \geq \chi_{0.05}^2(3) = 7.8147$, p 值 = 0, 拒绝 H_0 , 有显著的相关, 如图 14-5 所示。

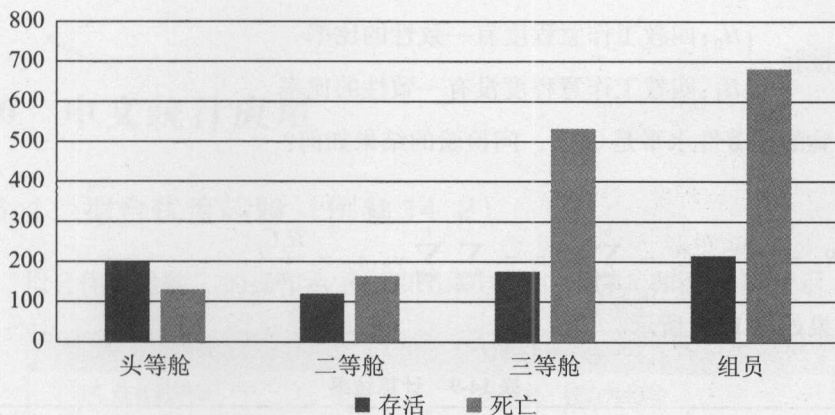


图 14-5 泰坦尼克号旅客身份和生死相关的条状图 (显著相关, 不是独立的)

(2) 性别和存活率的独立性检验:

H_0 : 身份和存活率是独立的

$$k^* = \sum_{i=1}^a \sum_{j=1}^b \frac{y_{ij}^2}{e_{ij}} - n = \frac{2223 (332)^2}{(1692)(706)} + \frac{2223 (374)^2}{(531)(706)} + \frac{2223 (1360)^2}{(1692)(1517)} \\ + \frac{2223 (157)^2}{(531)(1517)} - 2223 = 481.467$$

$k^* = 481.467 \geq \chi_{0.05}^2(1) = 3.8415$, p 值 = 0, 拒绝 H_0 , 有显著的相关。

例题 14.5 以下是某一大学, 教员 (分成: 教授, 副教授, 助理教授, 讲师 4 类) 对工作满意的程度 (分成: 满意, 普通, 不满意 3 种)。如图 14-6 所示。

		分 类 B				
分 类 A		讲师	助理教授	副教授	教授	总和
	满意	40	60	52	63	215
	普通	78	87	82	88	335
	不满意	57	63	66	64	250
	总和	175	210	200	215	800

图 14-6 计算教员对工作满意的程度

独立性检验： $\begin{cases} H_0: \text{教员与工作意程度是独立的} \\ H_1: \text{教员与工作意程度不是独立的} \end{cases}$

同构性检验： $\begin{cases} H_0: \text{四教工作意程度有一致性的比率} \\ H_1: \text{四教工作意程度没有一致性的比率} \end{cases}$

如果检验的显著性水平是 0.05，问检验的结果如何？

解答：

计算： $R_i = \sum_{j=1}^b y_{ij}, C_j = \sum_{i=1}^a y_{ij}, n = \sum_{i=1}^a \sum_{j=1}^b y_{ij}, e_{ij} = \frac{R_i C_j}{n}$

计算结果如表 14-9 所示。

表 14-9 计算结果

(i, j)	y_{ij}	e_{ij}	y_{ij}^2/e_{ij}	(i, j)	y_{ij}	e_{ij}	y_{ij}^2/e_{ij}
(1, 1)	40	47.03	34.0208	(1, 2)	60	56.44	63.7846
(1, 3)	52	53.75	50.3070	(1, 4)	63	57.78	68.6916
(2, 1)	48	73.28	83.0240	(2, 2)	87	87.94	86.0700
(2, 3)	82	83.75	80.2866	(2, 4)	88	90.03	86.0518
(3, 1)	57	54.69	59.4076	(3, 2)	63	65.62	60.4846
(3, 3)	66	62.50	69.6960	(3, 4)	64	67.19	60.9615

$$k^* = \sum_{i=1}^a \sum_{j=1}^b \frac{y_{ij}^2}{e_{ij}} - n = 802.7501 - 800 = 2.75$$

因为计算和独立性检验相同， $k^* = 2.75 < \chi^2_{0.05}(6) = 12.592$ ，所以接受 H_0 ，如图 14-7 所示， p 值 = 0.8394。

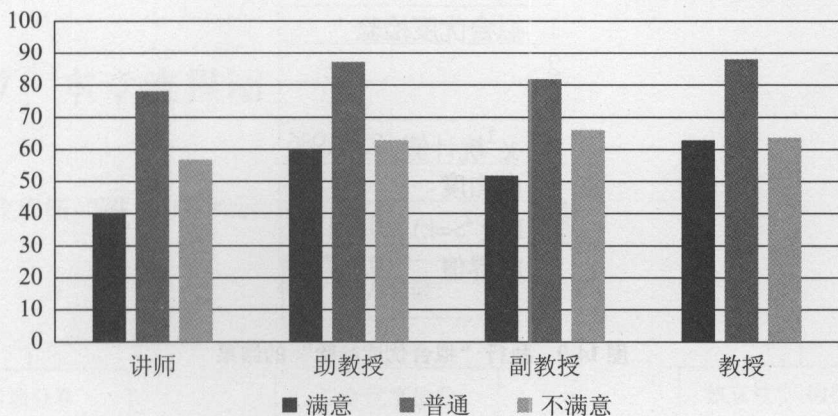


图 14-7 教员与工作满意度的条状图 (不显著相关, 独立)

14.6 中文统计应用

14.6.1 拟合优度检验 (例题 14.2)

执行“拟合优度检验”的操作示意图和结果分别如图 14-8 和图 14-9 所示。

13. 分类数据分析

14. 非参数统计

中文统计使用说明

陈文贤, 陈静枝 著《大话统计学》

拟和优度检验

列联表 - 独立性检验

列联表 - 独立性检验 (原始数据)

卡方检验 - 正态分布检验

卡方检验 - 中位数检验

A	B	C
1	12.3	
2	x	P(x)
3	0	0.1353
4	1	0.2707
5	2	0.2707
6	3	0.1804
7	4	0.0902
8	>4	0.0527
9		
10		
11		

拟合优度检验

输入

拟合优度的假设概率: B3:B8

观察值频数: C3:C9

估计参数数目: 1 ☐ 标志位于第一行

显著性水平: .05

输出选项

☐ 输出区域:

☒ 新工作表:

图 14-8 执行“拟合优度检验”的操作示意图

1	拟合优度检验	
2		
3		
4	χ^2 统计值	2.351986
5	自由度	4
6	$P(\chi^2 \geq k)$	0.671321
7	临界值	9.487729

图 14-9 执行“拟合优度检验”的结果

14.6.2 独立性检验（例题 14.4）

执行“独立性检验”的操作示意图和结果分别如图 14-10 和图 14-11 所示。

13. 分类数据分析	拟和优度检验
14. 非参数统计	列联表 - 独立性检验
中文统计使用说明	列联表 - 独立性检验 (原始数据)
陈文贤, 陈静枝 著《大话统计学》	卡方检验 - 正态分布检验
20	卡方检验 - 中位数检验

图 14-10 执行“独立性检验”的操作示意图

	A	B	C	D	E	F
1	列联表 - 独立性检验					
2						
3		头等	二等	三等	组员	总和
4	活	199	119	174	214	706
5	死	130	166	536	685	1517
6	总和	329	285	710	899	2223
7						
8	χ^2 值	181.8935				
9	自由度	3				
10	p值	0				
11	临界值	7.8147				

图 14-10 执行“独立性检验”的结果

14.7 本章流程图

本章流程图如图 14-12 所示。

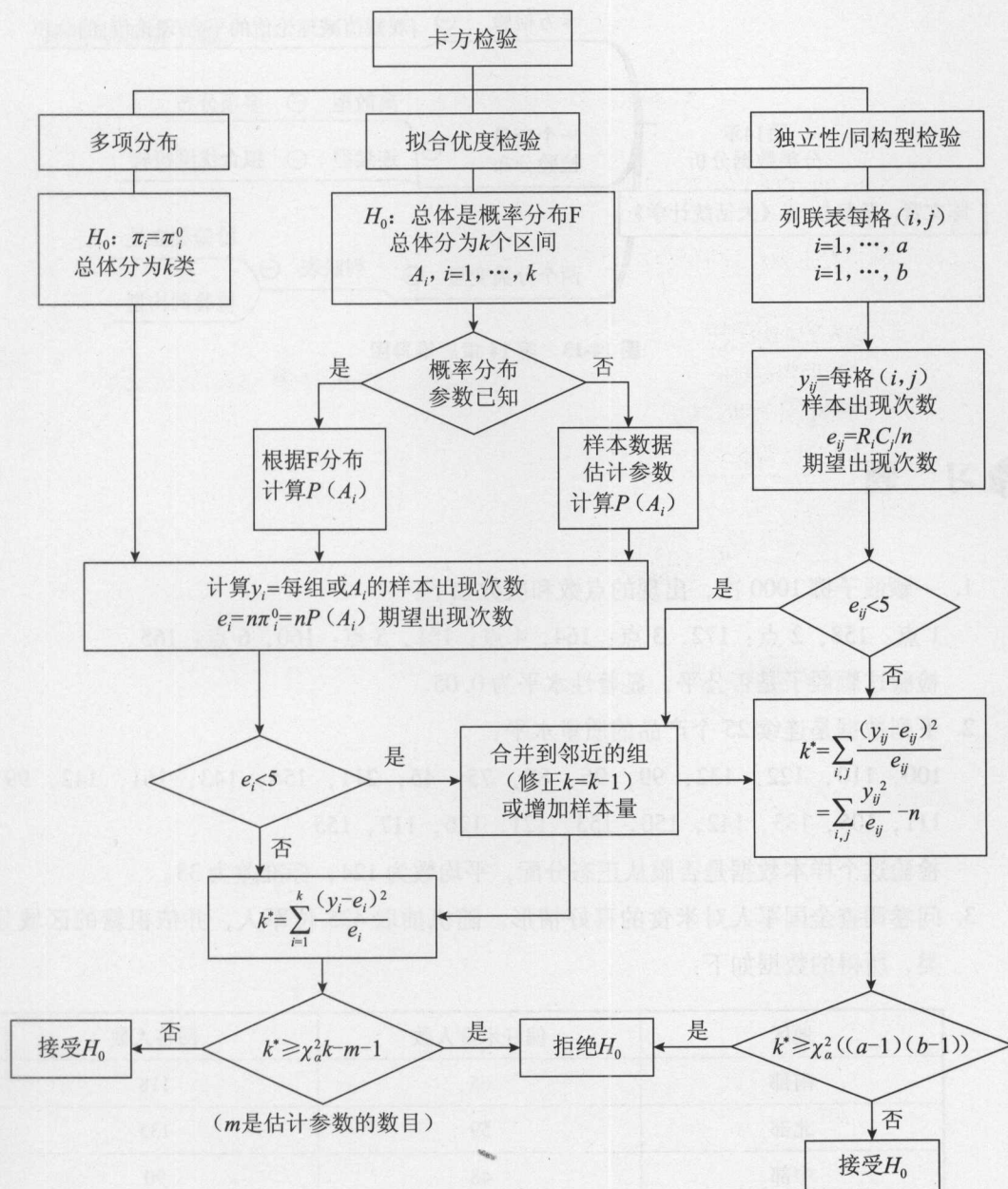


图 14-12 第 14 章流程图

14.8 本章思维导图

本章思维导图如图 14-13 所示。

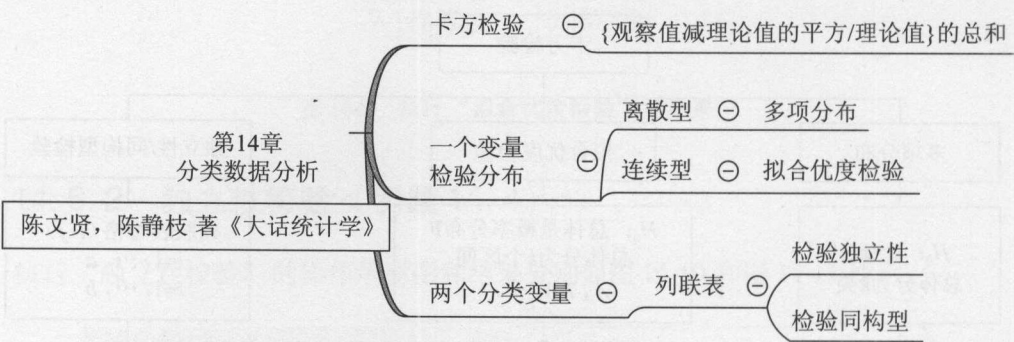


图 14-13 第 14 章思维导图

习 题

- 1. 一颗骰子掷 1000 次，出现的点数和次数如下：
1 点：158，2 点：172，3 点：164，4 点：181，5 点：160，6 点：165。
检验这颗骰子是否公平，显著性水平为 0.05。
- 2. 下列数据是连续 25 个产品的质量水平：
100，110，122，132，99，96，88，75，45，211，154，143，161，142，99，
111，105，133，142，150，153，121，126，117，155
检验这个样本数据是否服从正态分配，平均数为 124，标准差为 33。
- 3. 问卷调查全国军人对米食的喜好情形。随机抽取 435 位军人，并依祖籍的区域分类，所得的数据如下：

地区	偏好米食人数	回答人数
南部	65	118
北部	59	135
中部	48	90
西部	43	92

以 $\alpha = 0.05$ ，偏好米食的习惯，不同地区是否有显著的差异？

4. 心理学家测验白领阶级和蓝领阶级的工作态度，4 个蓝领阶级的分数：23，18，22，21。5 个白领阶级的分数：23，28，25，24，26。检验这两个阶级的工作态度有没有差异，显著性水平 0.05。

其他习题请下载。



第 15 章

非参数统计分析

无、名天地之始；有、名天地之母。故常无，欲以观其妙；常有，欲以观其徼。有之以为利，无之以为用。

——老子《道德经》

见山是山，见水是水。

——宋《五灯会元》

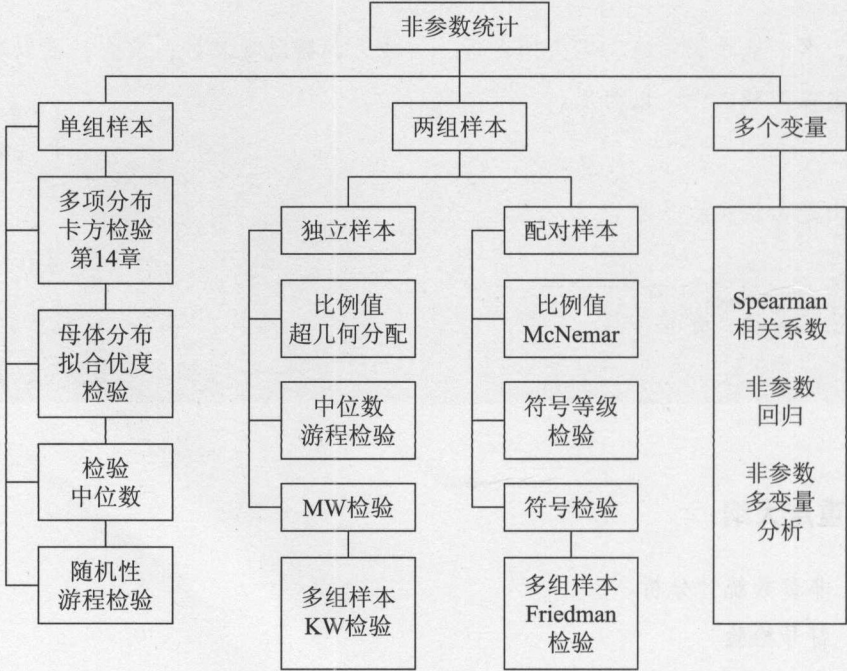
苟汰其芜杂，存其菁英

——《四库全书》（经部卷六）



本章重点大纲：

- 15.1 非参数统计分析
- 15.2 符号检验
- 15.3 符号秩检验
- 15.4 游程检验，检验随机性
- 15.5 Mann - Whitney 检验
- 15.6 Kruskal - Wallis 检验
- 15.7 Friedman 检验
- 15.8 Spearman 秩相关系数
- 15.9 中文统计应用
- 15.10 本章流程图
- 15.11 本章思维导图



本章概念图

15.1 非参数统计分析

在前面数章：假设检验，方差分析，回归分析等有关的检验中，都假定总体的分布是正态分布。但是如果这个假定稍微不符（总体的分布相差正态不远），或者样本量很大，上述检验的结果，仍然可以信赖，这种特性称作稳健性或鲁棒性（Robust）。如果总体分布，相差正态分布太多且样本量不大，则以上的检验就不适用。

有些统计推论不需要总体是正态分布的假定，这种推论我们称之为，无分布方法（distribution-free methods）。例如：游程检验（run test）和符号检验（sign test）是无分布方法。

有些统计推断与参数（parameter，例如： μ, σ, p ）无关，这种推论我们称之为，无参数方法（nonparametric methods）。例如：卡方独立性检验和游程检验是无参数方法。

我们定义“非参数检验”（nonparametric testing）包括“无分布方法”与“无参数方法”。所以，卡方检验的独立性检验，也算非参数统计，在前 14 章介绍。参数统计和非参数统计方法的比较如表 15-1 所示。

表 15-1 参数统计和非参数统计方法的比较

比较	有参数统计	非参数统计
目的	估计检验：平均数、变数、比例值、回归系数、相关系数	检验（没有区间估计）：中位数、母体分配、适合度、随机性、独立性
假设条件	正态母体或样本数大于 30	非正态母体且样本数小于 30
数据尺度	比率、区间尺度（平均数）、名目尺度（比例值）；很少有顺序尺度	以顺序尺度为主、名目尺度（连检验）；比率、区间尺度（降至顺序尺度）
特点	1. 计算多，区间尺度没浪费 2. 假设条件多，检验力较佳 3. 数据可以加减：总和、差 4. 标准概率分布： z, t, χ^2, F 5. 有回归、完整多因子 ANOVA	1. 顺序尺度，管理问题的问卷 2. 适合小样本 3. 统计量概率分布，有不同查表、 4. 函数式、仿真或用正态近似 5. 对极端值不受影响

“非参数检验”比“参数检验”有下列优点。

- (1) 假定条件较少，不需要对总体有太多的假定条件。
- (2) 计算较少，通常用比较数据的秩，没有太多的计算。
- (3) 适合小样本之研究。
- (4) 可以处理计数值的间断型数据，即分类与顺序尺度数据。

适合分类尺度数据的非参数统计有：卡方检验，游程检验等；适合顺序尺度数据的非参数统计有：符号检验，符号秩检验，Kolmogorov 检验，Mann - Whitney（MW）检验，KW 检验等。

非参数检验有下列缺点。

- (1) 检验力较（参数统计）差。
- (2) 将数据（从区间尺度）转换为顺序尺度，浪费数据的尺度与收集。
- (3) 每种检验法的统计量都要不同的概率表（大样本才能近似正态），缺乏通用的概率表（如 z 表，t 表）。本书将部分检验利用递归方程式导出其统计量的分布，但是要利用计算机计算。
- (4) 尚无可以检验有交互影响的二因子方差分析的非参数方法。

非参数统计检验总体的分布，多以“中位数”为主要参数，而参数统计是以“平均数”为主要参数。到底平均数估计和中位数估计的差别如何？兹说明如下。

(1) 中位数有一个好处，是它不受极端值的影响，譬如有一组样本数据，记录者或受访者将其中之一数据写成 10 倍，中位数可能不受影响，平均数就影响很大。如果随机抽样的样本包含有极大值或极小值，不管是否由于非抽样误差的数据收集错误，或实际存在差异于总体，检验中位数可能比检验平均数来得恰当。换言之，如果总体的分布有长而厚的尾巴，很有可能出现极大值或极小值，则中位数估计或检验比较合适。

(2) 平均数估计值可以用来估计总和。例如：估计每位推销员的平均销售量，乘以推销员的人数，则可以得到总销售量的估计。但是中位数估计值不可能用来估计总体的总和。不过，如果总体是顺序尺度或区间尺度，则总体总和的估计是无意义的（例如：每月总温度）。所以若是只要找出总体的中心点，不必用来估计总体的总和，而且总体不是正态分布，则利用中位数估计值。

检验总体分布有 3 个因素：位置（location）、分散（spread）、形态（shap），位置是集中趋势，分散是离差量数，形态是形态量数，如果原假设是检验总体分布（单总体或多总体），拒绝 H_0 表示 3 个因素至少有一个不相等。表 15-2 的因果关系、独立或相关，不是检验的目的或结果，是默认的样本设计（抽样或实验设计）。统计学无法确认因果关系，何者为因，何者为果，由研究者诠释。

表 15-2 统计检验的因果关系

统计方法		因果关系 样本独立 或相关	一个变量（果）				多个 变量
			单母体 无因果 关系	另一个分类的独立变量（因）			
				独立 双母体	独立 多母体	相依 双样本	相依 多样本
有 母 数 统 计	检验 平均数	z 检验 t 检验	z 检验 t 检验	ANOVA	t 检验	随机集区 设计	回归相关 分析 多因子 ANOVA 多变量 分析
	二项 比例值	z 检验	z 检验	χ^2 检验	McNemar test		
	检验 变异数	χ^2 检验	F 检验	Hartley 检验 Bartlett 检验			

续表

统计方法		因果关系 样本独立 或相关	一个变量（果）				多个 变量	
			单母体 无因果 关系	另一个分类的独立变量（因）				
				独立 双母体	独立 多母体	相依 双样本		相依 多样本
无 母 数 统 计	检验 中位数	符号检验 符号等级 检验	χ^2 检验 游程检验 MW U 检验	KW 检验	符号检验 符号等级	Friedman 检验	Spearman 相关系数	
	二项 比例值	二项 F 检验	超几何					
	多项 比例值	卡方	卡方 中位数					
	随机性	连检验						
	母体分配 Location Spread Shap	卡方 Kolmogorov 检验 Lilliefors 检验	MW U 检验					

参数统计和非参数统计的问题和方法的比较，如图 15-1 所示。

15.2 符号检验

符号检验（sign test），是检验总体中位数是否等于某假定值。如果是两组相依样本，符号检验可以检验其中位数是否相等。

15.2.1 检验单总体中位数

1. 符号检验的假定条件
- 1) X_1, X_2, \cdots, X_n 为总体随机变量的随机抽样。
- 2) 定义总体中位数为 M , $F(M) = P(X_i \leq M) = 0.5$ 。 M_0 为假定的中位数。
- 3) 检验:

$$\begin{cases} H_0^I : M = M_0 \\ H_1^I : M \neq M_0 \end{cases} \quad \text{左} \begin{cases} H_0^{II} : M \geq M_0 \\ H_1^{II} : M < M_0 \end{cases} \quad \text{右} \begin{cases} H_0^{III} : M \leq M_0 \\ H_1^{III} : M > M_0 \end{cases}$$

参数统计	单总体样本	检验均值 μ	z检验, t检验
		检验比例 π	z, 二项F分布, χ^2 检验
		检验方差 σ^2	χ^2 分布
	双总体独立样本	检验均值差 $\mu_1-\mu_2$	z检验, t检验
		检验比例差 $\pi_1-\pi_2$	z检验
		检验方差比 σ_1^2/σ_2^2	F 检验
	双总体配对样本	检验均值差 $\mu_1-\mu_2$	t检验
	多总体独立样本	检验均值全相等	单因素方差分析
		检验方差全相等	Bartlett检验
	多总体配对样本	检验均值全相等	二因素方差分析, 无重复
非参数统计	两组以上变量	回归与相关分析	一元回归
		多变量分析	多元回归
	单总体样本	卡方检验	多项分布
			拟合优度
			列联表独立性
		检验中位数	符号检验
		检验随机性	符号秩检验
			游程检验
		检验总体分布	Kolmogorov检验
		置信区间估计	Lilliefors检验
自力法 重迭法			
双总体独立样本		检验比例相等	超几何分布
	检验中位数相等	中位数卡方或游程检验	
	检验相同分布	Mann-Whitney检验	
	双总体配对样本	检验中位数相等	符号或符号秩检验
		多总体独立样本	检验相同分布
	多总体配对样本		检验相同分布
	两组以上变量	相关分析	Spearman秩相关系数
		回归分析	Brown-Mood法或Theil法
	时间序列, 指数		

图 15-1 参数统计和非参数统计

4) 假如 p 为总体数据中小于 M 的比例。所以上述检验相当于比例检验:

$$\begin{cases} H_0^I: p = 0.5 \\ H_1^I: p \neq 0.5 \end{cases} \quad \text{左} \begin{cases} H_0^{II}: p \geq 0.5 \\ H_1^{II}: p < 0.5 \end{cases} \quad \text{右} \begin{cases} H_0^{III}: p \leq 0.5 \\ H_1^{III}: p > 0.5 \end{cases}$$

2. 检验步骤

1) 如果样本数据 $\{x_1, x_2, \dots, x_n\}$ 中有等于 M_0 的数据, 则样本删除这些等于 M_0 的数据。样本量 n 改为新的样本量。

2) 计算 $x_-^* =$ 所有样本数据 $\{x_1, x_2, \dots, x_n\}$ 中小于 M_0 的数目。

计算 $x_+^* =$ 所有样本数据 $\{x_1, x_2, \dots, x_n\}$ 中大于 M_0 的数目。

3) 检验值 $v^* = \min\{x_-^*, x_+^*\} =$ 双侧检验 H_0^I 的统计值。

$v^* = x_-^* =$ 左侧检验 H_0^{II} 的统计值。

$v^* = x_+^* =$ 右侧检验 H_0^{III} 的统计值。

4) 利用二项分布

$$P\{B(n, 0.5) \leq v^*\} = \sum_{i=0}^{v^*} \binom{n}{i} (0.5)^n$$

双侧检验 H_0^I , p 值 $= 2P\{B(n, 0.5) \leq v^*\}$ 。

左侧检验 H_0^{II} , p 值 $= P\{B(n, 0.5) \leq v^*\}$ 。

右侧检验 H_0^{III} , p 值 $= P\{B(n, 0.5) \geq v^*\}$ 。

如果 p 值 $\leq \alpha$, 则拒绝 H_0 。

5) 利用正态分布近似, 若 $n \geq 10$, 则 $B(n, 0.5)$ 近似正态分布 $N(0.5n, 0.25n)$:

$$z^* = \frac{2v^* - n}{\sqrt{n}}$$

$$\text{或 连续性修正} \begin{cases} z^* = \frac{(v^* + 0.5) - 0.5n}{0.5\sqrt{n}} & \text{if } v^* < 0.5n \\ z^* = \frac{(v^* - 0.5) - 0.5n}{0.5\sqrt{n}} & \text{if } v^* > 0.5n \end{cases}$$

双侧检验 H_0^I , p 值 $= 2P\{Z \geq |z^*|\}$ 。

左侧检验 H_0^{II} , p 值 $= P\{Z \leq z^*\}$ 。

右侧检验 H_0^{III} , p 值 $= P\{Z \geq z^*\}$ 。

如果 p 值 $\leq \alpha$, 则拒绝 H_0 。

例题 15.1 某大医院抽样 20 个病人, 挂号及等候的时间 (分钟) 如下:

22, 30, 31, 40, 37, 25, 29, 14, 30, 17, 23, 32, 20, 40, 28, 26, 33, 25,

34, 21

假设中位数为 M 。

检验:
$$\begin{cases} H_0: M \leq 25 \\ H_1: M > 25 \end{cases}$$

如果检验的显著性水平是 0.05，问检验的结果如何？

解答：20 个数据中有 2 个等于 25，所以去掉。 $n = 18$ 。

$$x_-^* = 6, x_+^* = 18 - 6 = 12, v^* = 12$$

$$p \text{ 值} = P\{B(n, 0.5) \geq v^*\} = P\{B(18, 0.5) \geq 12\} = 1 - P\{B(18, 0.5) \leq 11\}$$
$$= 1 - 0.881 = 0.119$$

检验结果：接受 H_0 ，中位数小于 25。

利用近似正态分布：

$$z^* = \frac{2v^* - n}{\sqrt{n}} = \frac{2 \times 12 - 18}{\sqrt{18}} = 1.4142$$
$$P(Z \geq 1.4142) = 0.079$$

连续性修正：

$$z^* = \frac{(v^* - 0.5) - 0.5n}{0.5\sqrt{n}} = \frac{(12 - 0.5) - 0.5 \times 18}{0.5\sqrt{18}} = 1.1735$$
$$P(Z \geq 1.1735) = 0.1203$$

经过连续性修正， p 值较近似。

符号检验步骤整理如表 15-3 所示。

表 15-3 符号检验步骤

检验方法	双侧检验 $H_0: M = M_0$ $H_1: M \neq M_0$	左侧检验 $H_0: M \geq M_0$ $H_1: M < M_0$	右侧检验 $H_0: M \leq M_0$ $H_1: M > M_0$
二项分布	$v^* = \min\{x_-^*, x_+^*\}$ $p \text{ 值} = 2P\{B(n, 0.5) \leq v^*\}$ 若 $p \text{ 值} < \alpha$ ，则拒绝 H_0	$v^* = x_+^*$ $p \text{ 值} = P\{B(n, 0.5) \leq v^*\}$ 若 $p \text{ 值} < \alpha$ ，则拒绝 H_0	$v^* = x_+^*$ $p \text{ 值} = P\{B(n, 0.5) \geq v^*\}$ 若 $p \text{ 值} < \alpha$ ，则拒绝 H_0
正态分布 (近似)	z^* 定义在公式 (15.1) $p \text{ 值} = 2P\{Z \geq z^* \}$ 若 $p \text{ 值} < \alpha$ ，则拒绝 H_0	z^* 定义在公式 (15.1) $p \text{ 值} = P\{Z \leq z^*\}$ 若 $p \text{ 值} < \alpha$ ，则拒绝 H_0	z^* 定义在公式 (15.1) $p \text{ 值} = P\{Z \geq z^*\}$ 若 $p \text{ 值} < \alpha$ ，则拒绝 H_0

15.2.2 检验双总体成对样本

两组成对或相依样本，符号检验可以检验其中位数是否相等。

检验步骤:

- 1) 如果样本数据 $\{x_1, x_2, \dots, x_n\}$ 与样本数据 $\{y_1, y_2, \dots, y_n\}$ 是“成对/相依性”样本。
- 2) 计算 $d_i = x_i - y_i$, 如果样本数据 $\{d_1, d_2, \dots, d_n\}$ 中有等于 0 的数据, 则样本删除这些等于 $d_i = 0$ 的数据。样本量 n 改为新的样本量。
- 3) 计算 $x_-^* =$ 所有数据 $\{d_1, d_2, \dots, d_n\}$ 中小于 0 的数目。
计算 $x_+^* =$ 所有样本数据 $\{d_1, d_2, \dots, d_n\}$ 中大于 0 的数目。
- 4) 令 $M_0 = 0$, 以下步骤同符号检验步骤的第 4 步。

例题 15.2 (见网络资源)

15.3 符号秩检验

符号秩检验 (sign rank test, 又称为 Wilcoxon test), 检验总体中位数, 而其分布有对称性。虽然要求条件比符号检验多一个对称性, 但是其检验功效较“符号检验”更强。所以如果知道总体分布有对称性, 则应该采用 Wilcoxon test。但是比起检验平均数的“t 检验”, 符号秩检验的检验功效稍差, 不过 t 检验需要总体是正态分布的假定。如果是两组相依样本, 符号秩 Wilcoxon 检验可以检验其中位数是否相等。

15.3.1 检验单总体中位数

1. 符号秩 Wilcoxon 检验的假定条件

1) X_1, X_2, \dots, X_n 为总体随机变量的随机抽样, 其分布是对称型 (symmetric), 累积概率分布为 F 。

2) 定义总体中位数为 M , $F(M) = P(X_i \leq M) = 0.5$, M_0 为假定的中位数。

3) 检验:

$$\begin{cases} H_0^I : M = M_0 \\ H_1^I : M \neq M_0 \end{cases} \quad \text{左} \begin{cases} H_0^{II} : M \geq M_0 \\ H_1^{II} : M < M_0 \end{cases} \quad \text{右} \begin{cases} H_0^{III} : M \leq M_0 \\ H_1^{III} : M > M_0 \end{cases}$$

2. 检验步骤

- 1) 令 $y_i = x_i - M_0 =$ 所有样本数据 $\{x_1, x_2, \dots, x_n\}$ 减去 M_0 。
- 2) 如果有 $y_i = 0$, 则不计算顺序, 样本量减 1。
- 3) 将 y_i 取绝对值 $|y_i|$, 然后按大小, 由小到大排列顺序。绝对值最小的秩或等级 (rank) 为 1, 最大者的顺序为 n 。如果绝对值相等, 顺序取其平均数, 例如两个绝对值相

等，则两个的顺序为原有顺序相加除以 2。

4) 统计量 $X_-^* =$ 所有随机抽样 $\{X_1, X_2, \dots, X_n\}$ 中，小于 M_0 的顺序加起来。

计算统计值 $x_-^* =$ 所有样本数据 $\{x_1, x_2, \dots, x_n\}$ 中，小于 M_0 的顺序加起来。

统计量 $X_+^* =$ 所有随机抽样 $\{X_1, X_2, \dots, X_n\}$ 中，大于 M_0 的顺序加起来。

计算统计值 $x_+^* =$ 所有样本数据 $\{x_1, x_2, \dots, x_n\}$ 中，大于 M_0 的顺序加起来。

5) $V = \min\{X_-^*, X_+^*\} =$ 双侧检验的统计量。

$v^* = \min\{x_-^*, x_+^*\} =$ 双侧检验的统计值。

$V = X_+^* =$ 左侧与右侧检验的统计量。

$v^* = x_+^* =$ 左侧与右侧检验的统计值。

6) 双侧检验 H_0^I ， p 值 $= 2P\{V \leq v^*\}$ 。

左侧检验 H_0^{II} ， p 值 $= P\{V \leq v^*\}$ 。

右侧检验 H_0^{III} ， p 值 $= P\{V \geq v^*\}$ 。

若 p 值 $\leq \alpha$ ，则拒绝 H_0 。

7) V 的分布可利用递归方程式 (recursive equations) 求得。请见补充教材。

8) 利用表 A7，得 T_L, T_U 。双侧检验，若 $v^* \leq T_L$ 或 $v^* \geq T_U$ ，则拒绝 H_0 ；左侧检验，若 $v^* \leq T_L$ ，则拒绝 H_0 ；右侧检验，若 $v^* \geq T_U$ ，则拒绝 H_0 。

9. 利用正态分布 (近似)。当 $n \geq 15$ ，则 V 近似正态分布，平均数 μ ，方差 σ^2 ，即

$$V \sim N(\mu, \sigma^2), \mu = \frac{n(n+1)}{4}, \sigma^2 = \frac{n(n+1)(2n+1)}{24}$$
$$z^* = \frac{x_+^* - \mu}{\sigma} \tag{15-2}$$

双侧检验 H_0^I ， p 值 $= 2P\{Z \geq |z^*|\}$ 。左侧检验 H_0^{II} ， p 值 $= P\{Z \leq z^*\}$ 。

右侧检验 H_0^{III} ， p 值 $= P\{Z \geq z^*\}$ 。

若 p 值 $\leq \alpha$ ，则拒绝 H_0 。

符号秩 Wilcoxon 检验步骤整理如表 15-4 所示。

表 15-4 符号秩 Wilcoxon 检验步骤

检验方法	双侧检验 $H_0:M = M_0$ $H_1:M \neq M_0$	左侧检验 $H_0:M \geq M_0$ $H_1:M < M_0$	右侧检验 $H_0:M \leq M_0$ $H_1:M > M_0$
递归方程式	补充教材	补充教材	补充教材

续表

正态分布 (近似)	$z^* = \frac{x_+^* - \mu}{\sigma} \text{ 公式 (15-2)}$ $p \text{ 值} = 2P\{Z \geq z^* \}$ 若 $p \text{ 值} < \alpha$, 则拒绝 H_0	$z^* = \frac{x_+^* - \mu}{\sigma}$ $p \text{ 值} = P\{Z \leq z^*\}$ 若 $p \text{ 值} < \alpha$, 则拒绝 H_0	$z^* = \frac{x_+^* - \mu}{\sigma}$ $p \text{ 值} = P\{Z \geq z^*\}$ 若 $p \text{ 值} < \alpha$, 则拒绝 H_0
查表法	$t^* = \min\{x_-^*, x_+^*\}$ 查表 A-7, 得 T_L, T_U 。 若 $\nu^* \leq T_L$ 或 $\nu^* \geq T_U$, 则拒绝 H_0	$t^* = \min\{x_-^*, x_+^*\}$ 查表 A-7, 得 T_L 若 $\nu^* \leq T_L$, 则拒绝 H_0	$t^* = \max\{x_-^*, x_+^*\}$ 查表 A-7, 得 T_U 若 $\nu^* \geq T_U$, 则拒绝 H_0

例题 15.3 (见网络资源)

15.3.2 检验双总体成对样本

两组成对/相依样本, 符号秩检验可以检验其中位数是否相等, 检验步骤:

- 1) 如果样本数据 $\{x_1, x_2, \dots, x_n\}$ 与样本数据 $\{y_1, y_2, \dots, y_n\}$ 是“成对/相依性”样本。
- 2) 计算 $d_i = x_i - y_i$, 如果样本数据 $\{d_1, d_2, \dots, d_n\}$ 中有等于 0 的数据, 则样本删除这些等于 $d_i = 0$ 的数据。样本量 n 改为新的样本量。
- 3) 计算 x_-^* = 所有数据 $\{x_1, x_2, \dots, x_n\}$ 中, 小于 0 的数目。
 计算 x_+^* = 所有样本数据 $\{x_1, x_2, \dots, x_n\}$ 中, 大于 0 的数目。
- 4) 令 $M_0 = 0$, 以下步骤同符号秩 Wilcoxon 检验步骤的第 5 步。

15.4 游程检验, 检验随机性

游程检验 (run test), 是检验样本数据是否有随机性, 因为在统计检验, 不管是参数检验或非参数统计, 通常需要样本数据具有随机性。游程检验也可以检验两组独立样本, 是否有相同的分布。

15.4.1 检验单总体样本数据的随机性

1. 游程检验的假定条件

- 1) X_1, X_2, \dots, X_n 为总体随机变量的随机抽样。其数据出现是有顺序的。即 X_i 出现在 X_{i-1} 之后, 在 X_{i+1} 之前。
- 2) X_i 数据可以是二分类的分类尺度, 也可以是顺序尺度, 区间尺度或比率尺度。

- 3) 检验: $\begin{cases} H_0: \text{样本数据具有随机性} \\ H_1: \text{样本数据不具有随机性} \end{cases}$

4) 定义数据中连续的 1 或 0 为一个游程 (run)。即连续的 1 前后为 0, 则为一个游程。如果只有一个 1 前后为 0, 也算一个游程。

例如: A BBB A BBB AA BB A B AA, 是 9 个游程。

++++- - - - + + + - - - + +, 16 个数据有 5 个游程。

5) 如果“游程”的数目太少或太多, 表示数据有规则性, 不具有随机性。

例如: - - - - - - - + + + + + + + +, 16 个数据只有 2 个游程。

+ - + - + - + - + - + - + - + -, 16 个数据有 16 个游程。

检验法则: 游程的数目在一个区间 $[L, U]$, 则接受 H_0 。

2. 检验步骤

1) 将数据转换为 0 或 1; 或是二分类的符号, 例如: 正负号 (+, -)。可以将数据减去平均数或中位数后, 小于等于 0 者为 0 (或 -), 大于 0 者为 1 (或 +)。

2) 令 m 为 0 (或 -) 的个数, n 为 1 (或 +) 的个数。

3) R = 随机变量的游程的数目, 为检验 H_0 的统计量。

r^* = 样本数据的游程的数目, 为检验 H_0 的统计值。

4) 利用 R 的分布, R 的分布如下:

$$P\{R = 2k\} = 2 \frac{\binom{m-1}{k-1} \binom{n-1}{k-1}}{\binom{m+n}{k}} \quad P\{R = 2k+1\} = \frac{\left[\binom{m-1}{k-1} \binom{n-1}{k} + \binom{m-1}{k} \binom{n-1}{k-1} \right]}{\binom{m+n}{k}}$$

p 值 = $2\min(P\{R \leq r^*\}, P\{R \geq r^*\})$ 。若 p 值 $\leq \alpha$, 则拒绝 H_0 。

5) 查表 A-10 得 L_r, U_r 。若 $r^* \leq L_r$ 或 $r^* \geq U_r$, 则拒绝 H_0 。

6) 查表 A-11 得 $P(R \leq r^*)$ 。

p 值 = $2\min(P\{R \leq r^*\}, P\{R \geq r^*\})$

p 值 = $2\min(P\{R \leq r^*\}, 1 - P\{R \leq r^*\})$ 。

若 p 值 $\leq \alpha$, 则拒绝 H_0 。

7) 利用正态分布, 当 $n \geq 10, m \geq 10$, 则 R 近似正态分布, 平均数为 μ , 方差为 σ^2 。

$$R \sim N(\mu, \sigma^2), \mu = \frac{2mn + n + m}{n + m}, \sigma^2 = \frac{2nm(2nm - n - m)}{(n + m)^2(n + m - 1)} \quad (15-4)$$

计算 $z^* = \frac{r^* - \mu}{\sigma}$, p 值 = $2P\{Z \geq |z^*|\}$, 若 p 值 $\leq \alpha$, 则拒绝 H_0 。

游程检验步骤整理如表 15-5 所示。

表 15-5 游程检验步骤

| 利用 R 的分布（用计算机） | 正态分布 | 查表法 |
|--|---|---|
| $r^* = \text{游程的数目}$
$P(R = i)$ 的分布公式 (15-3)
$p \text{ 值} = 2\min(P\{R \leq r^*\}, P\{R \geq r^*\})$ | $r^* = \text{游程的数目}$
$z^* = \frac{r^* - \mu}{\sigma}$
公式 (15-4) | $r^* = \text{游程的数目}$
查表 A-10 得 L_r, U_r
若 $r^* \leq L_r$ 或 $r^* \geq U_r$, 则拒绝 H_0
查表 A-11 得 $P(R \leq r^*)$
$p \text{ 值} = 2\min(P\{R \leq r^*\}, P\{R \geq r^*\})$ |

例题 15.4 （见网络资源）

15.4.2 检验双总体独立样本

游程检验检验两组独立样本，是否有相同的分布。

1. 假定条件

1) X_1, X_2, \dots, X_m 与 Y_1, Y_2, \dots, Y_n 为两组随机抽样，分别有总体累积概率分布 F_X 与 F_Y 。

2) 检验: $\begin{cases} H_0: F_X = F_Y \\ H_A: F_X \neq F_Y \end{cases}$

2. 检验步骤

1) 将两组样本数据混合，从小到大排列顺序。

2) 在排好顺序的数据上，若 x_i ，则记作 +；若 y_i ，则记作 -。

3) $r^* = \text{样本数据的游程的数目}$ ，为检验 H_0 的统计值。

4) 以下步骤同上述游程检验步骤的第 4 步。

15.5 Mann - Whitney 检验

双样本检验相同分布，是检验两个“独立样本”的分布是否相等。检验的方法有很多，在本书介绍的检验方法有：

1) 正态总体“t 检验”，在第 11.3 到 11.5 节介绍，参数统计。

2) 中位数（卡方）检验，在第 14.6 节介绍。

3) 游程检验，在第 15.4 节介绍。

4) 秩总和检验（Wilcoxon rank - sum test 或 Mann - Whitney U test），本节介绍。

秩总和检验的检验功效较前两者强，但是较 t 检验弱。秩总和检验的要求条件是两总

体的概率分布形状相同,即方差相同。秩总和检验以下称为 Mann - Whitney 检验 (因为 Wilcoxon 检验已经命名为符号秩检验)。

15.5.1 Mann - Whitney 检验

1. Mann - Whitney 检验的假定条件

1) X_1, X_2, \dots, X_n 与 Y_1, Y_2, \dots, Y_m 为两组独立随机抽样, $n \leq m$, 其概率分布分别是 F 与 G, 而且方差相等。

2) 检验:

$$\begin{cases} H_0^I: F = G, M_x - M_y = d_0 \\ H_1^I: F \neq G, M_x - M_y \neq d_0 \end{cases} \quad \text{左} \begin{cases} H_0^{II}: M_x - M_y \geq d_0 \\ H_1^{II}: M_x - M_y < d_0 \end{cases} \quad \text{右} \begin{cases} H_0^{III}: M_x - M_y \leq d_0 \\ H_1^{III}: M_x - M_y > d_0 \end{cases}$$

2. 检验步骤

1) 所有随机抽样 $\{X_1 - d_0, X_2 - d_0, \dots, X_n - d_0, Y_1, Y_2, \dots, Y_m\}$ 合并, 按大小顺序, 由小到大排列。所有样本数据 $\{x_1 - d_0, x_2 - d_0, \dots, x_n - d_0, y_1, y_2, \dots, y_m\}$ 合并, 按大小顺序, 由小到大排列。($m \geq n$)

2) 令 $R_i = X_i - d_0$ 的排列顺序 (随机变量): $T = \sum_{i=1}^n R_i$ 。

计算 $r_i = x_i - d_0$ 的排列顺序; $t^* = \sum_{i=1}^n r_i$ = 检验 $H_0^I, H_0^{II}, H_0^{III}$ 的统计值。

3) 双侧检验 H_0^I , p 值 = $2\min(P\{T \leq t^*\}, P\{T \geq t^*\})$ 。

p 值 = $2\min(P\{T \leq t^*\}, 1 - P\{T < t^*\}) = 2\min(P\{T \leq t^*\}, 1 - P\{T \leq t^* - 1\})$ 。

左侧检验 H_0^{II} , p 值 = $P\{T \leq t^*\}$ 。

右侧检验 H_0^{III} , p 值 = $P\{T \geq t^*\}$ 。

若 p 值 $\leq \alpha$, 则拒绝 H_0 。

4) T 的分布可利用递归方程式 (recursive equations) 求得, 请见附录光盘补充教材。

5) 秩总和检验 (Wilcoxon rank sum test): 查表 A-8, 得到 T_L, T_U

若 $T_L < t^* < T_U$, 则接受 H_0 。

6) MW U 分布 (MW U 检验): 查表 A-9, 得 u_α 。

计算 $u_1 = mn + \frac{n(n+1)}{2} - t^*, u_2 = mn + \frac{m(m+1)}{2} - s^*, u = \min\{u_1, u_2\}$ 。

若 $u \leq u_\alpha$, 则拒绝 H_0 。

7) 正态分布: 当 $n, m \geq 15$, 则 T 近似正态分布, 平均数为 μ , 方差为 σ^2 。

$$T \sim N(\mu, \sigma^2), \mu = \frac{n(n+m+1)}{2}, \sigma^2 = \frac{nm(n+m+1)}{12} \tag{15-5}$$

计算 $z^* = \frac{t^* - \mu}{\sigma}$ ，双侧检验 H_0^I ， p 值 = $2P\{Z \geq |z^*|\}$ 。

左侧检验 H_0^{II} ， p 值 = $P\{Z \leq z^*\}$ 。右侧检验 H_0^{III} ， p 值 = $P\{Z \geq z^*\}$ 。

若 p 值 $\leq \alpha$ ，则拒绝 H_0 。

8) 双侧检验 H_0^I ，可利用模拟的方法，计算 p 值，请见补充教材。

Mann - Whitney 检验步骤整理如表 15-6 所示。

表 15-6 Mann - Whitney 检验步骤

| 检验方法 | 双侧
$H_0:M_x = M_y$
$H_1:M_x \neq M_y$ | 左侧
$H_0:M_x \geq M_y$
$H_1:M_x < M_y$ | 右侧
$H_0:M_x \leq M_y$
$H_1:M_x > M_y$ |
|--------------|---|--|--|
| 递归方程式 | 补充教材 | 补充教材 | 补充教材 |
| 正态分布
(近似) | $z^* = \frac{t^* - \mu}{\sigma}$ 公式 (15.5)
p 值 = $2P\{Z \geq z^* \}$
若 p 值 $< \alpha$ ，则拒绝 H_0 | $z^* = \frac{t^* - \mu}{\sigma}$
p 值 = $P\{Z \leq z^*\}$
若 p 值 $< \alpha$ ，则拒绝 H_0 | $z^* = \frac{t^* - \mu}{\sigma}$
p 值 = $P\{Z \geq z^*\}$
若 p 值 $< \alpha$ ，则拒绝 H_0 |
| 查表法 | 秩总和检验查表 A-8
MW U 检验查表 A-9 | 秩总和检验查表 A-8
MW U 检验查表 A-9 | 秩总和检验查表 A-8
MW U 检验查表 A-9 |
| 模拟法 | 补充教材 | 补充教材 | 补充教材 |

例题 15.5 (见网络资源)

15.6 Kruskal-Wallis 检验

Kruskal - Wallis 检验 (简称 KW 检验) 是检验两个以上总体独立样本有相同的中位数或分布。它和方差分析的假定相同，但是不同的是 KW 检验不需要正态分布的假定条件。KW 检验是双样本秩总合检验的扩大到两组以上样本。

1. KW 检验的假定条件

- 1) 总体数据是连续型的分布。
- 2) k 组 (总体或处理) 的 “独立” 样本。
- 3) 每组样本的样本量是 $n_i, i = 1, \cdots, k$ 。
- 4) 总共的样本量 $N = \sum n_i$ 。

5) 样本数据是 $x_{ij}, i = 1, 2, \dots, k, j = 1, 2, \dots, n_i$ (i 是总体编号, j 是样本编号, 如图 15-2 所示)。

6) 检验的假设是 $H_0: M_1 = M_2 = \dots = M_k$ (k 个总体有相同的中位数或分布)。

| | | 样本 | | | | | |
|----|-----|----|---|-----|-----|-----|-----------------------|
| | | 1 | 2 | ... | j | ... | n_i |
| 总体 | 1 | | | | | | $\sum_j r_{1j} = R_1$ |
| | 2 | | | | | | |
| | ... | | | | | | |
| | i | | | | | | $\sum_j r_{ij} = R_i$ |
| | ... | | | | | | |
| | k | | | | | | $\sum_j r_{kj} = R_k$ |

图 15-2 样本数据

2. 检验步骤:

- 1) 将 x_{ij} “全部” 按照大小, 由小到大排列。
- 2) $r_{ij} = x_{ij}$ 在所有 N 个数据的排序, 1 是最小, N 是最大, 相同大小, 排序取平均。
- 3) $R_i = \sum_{j=1}^{n_i} r_{ij}, \bar{R}_i = \frac{R_i}{n_i}$ [R_i 是第 i 组 (总体) 样本数据之排序的总和]。
- 4) 检验值 $H = \frac{12}{N(N+1)} \sum_{i=1}^k n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2 = \frac{12}{N(N+1)} \sum_{i=1}^k \left(\frac{R_i^2}{n_i} \right) - 3(N+1)$ 。
- 5) 若 $H \geq \chi^2_{\alpha, (k-1)}$, 则拒绝 H_0 。

例题 15.6 (见网络资源)

15.7 Friedman 检验

Friedman 检验是检验两个以上总体的相依 (区组, 配对) 样本的平均数是否全部相等。它和方差分析的随机区组设计的假定相同, 但是不同的是 Friedman 检验不需要正态分布的假定条件。Friedman 检验是双样本符号检验的扩大到两组以上样本。

1. Friedman 检验的假定条件

- 1) 总体数据是连续型的分布。
- 2) k 组 (k 个总体或处理) 之 “相依” 样本。
- 3) 每组样本的样本量相等是 n 。
- 4) 总共的样本量 $N = \sum n = kn$ 。

5) 样本数据是 $x_{ij}, i = 1, 2, \dots, n, j = 1, 2, \dots, k$ (i 是集区编号, j 是总体编号, 如图 15-3 所示)。

6) 检验的假设是 $H_0: M_1 = M_2 = \dots = M_k$ 。

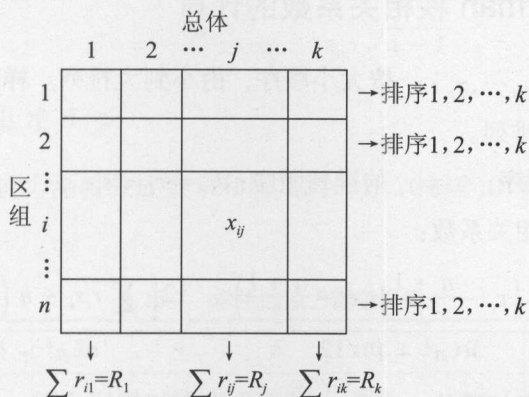


图 15-3 样本数据

2. 检验步骤

- 1) 将 x_{ij} 在“每个集区 i ”按照大小, 由小到大排列。
- 2) $R_{ij} = x_{ij}$ 在第 i 个集区数据的排序, 1 是最小, k 是最大, 相同大小, 排序取平均。
- 3) $R_j = \sum_{i=1}^n R_{ij}$ = 在第 j 组 (总体) 样本数据之排序的总和。
- 4) 检验值 $H = \frac{12}{nk(k+1)} \sum_{j=1}^k R_j^2 - 3n(k+1)$ 。
- 5) 若 $H \geq \chi_{\alpha}^2(k-1)$, 则拒绝 H_0 。

例题 15.7 (见网络资源)

15.8 Spearman 秩相关系数

Spearman 秩相关系数 (Spearman's rank correlation coefficient), r_{sp} 相对于相关分析的相关系数, 只是它不需要总体正态分布的假定条件, 同时它适用于顺序尺度的数据。

15.8.1 Spearman 秩相关系数的假设

Spearman 秩相关系数的假设: 变量 X 与变量 Y 无线性相关性。

- 1) X_1, X_2, \dots, X_n 与 Y_1, Y_2, \dots, Y_n 为两组相依随机抽样。

- 2) 检验: $\begin{cases} H_0: & \text{变量 } X \text{ 与变量 } Y \text{ 无线性相关性} \\ H_1: & \text{变量 } X \text{ 与变量 } Y \text{ 有线性相关性} \end{cases}$

15.8.2 Spearman 秩相关系数的计算

1) 样本数据 $\{x_1, x_2, \dots, x_n\}$, 按大小顺序, 由小到大排列。样本数据 $\{y_1, y_2, \dots, y_n\}$, 按大小顺序, 由小到大排列。

2) $r_i = x_i$ 的排列顺序; $s_i = y_i$ 的排列顺序; $d_i = r_i - s_i$ 。

3) 以 r_i 和 s_i 计算相关系数:

$$r_{sp} = \frac{S_{rs}}{\sqrt{S_{rr}S_{ss}}} = \frac{\sum \left(r_i - \frac{n+1}{2}\right) \left(s_i - \frac{n+1}{2}\right)}{n(n^2-1)/12} = \frac{12 \left[\sum r_i s_i - n \left(\frac{n+1}{2}\right)^2 \right]}{n(n^2-1)}$$

4) 如果 x_i 数据没有相等的, 而且 y_i 数据也没有相等的, 则

$$r_{sp} = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

5) 如果 x_i 没有相同的值且 y_i 没有相同的值, 则以上两个 r_{sp} 的公式相等。因为 x_i 没有相同的值且 y_i 没有相同的值, 则 $\sum d^2$ 的最小值是 0, 即 x_i 与 y_i 有相同的秩; $\sum d^2$ 的最大值是 $n(n^2-1)/3$, 即 x_i 与 y_i 有相反的秩。所谓 x_i 与 y_i 有相反的秩, 即 x_1 是最小, y_1 是最大; x_2 是第二小, y_2 是第二大; 以此类推, x_n 是最大, y_n 是最小。所以:

$$0 \leq \sum d^2 \leq \frac{n(n^2-1)}{3} \quad -1 \leq r_{sp} \leq 1$$

6) 如果 x_i 有相同的值或 y_i 有相同的值, 则以上两个 r_{sp} 的公式并不相等, 但是因为第二个公式计算比较简单, 所以常用第二个公式计算 Spearman 秩相关系数, 即

$$r_{sp} = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

15.8.3 检验 Spearman 秩相关系数

- 检验: $\begin{cases} H_0: \rho_s = 0, & \text{变量 } X \text{ 与变量 } Y \text{ 无线性相关性} \\ H_A: \rho_s \neq 0 & \text{变量 } X \text{ 与变量 } Y \text{ 有线性相关性} \end{cases}$

检验步骤:

1) 计算 r_{sp} 。

2) 查表 A-12 (A-12 见网络资源), 双侧检验 α , 得到 r_α 。若 $r_{sp} > r_\alpha$ 或 $r_{sp} < -r_\alpha$, 则拒绝 H_0 。

3) 利用正态分布: 若 $n \geq 10$, 则在原假设之下, r_{sp} 的分布近似正态分布, 平均数为 0, 标准差 $\sigma_{r_{sp}} = \frac{1}{\sqrt{n-1}}$ 。检验值 z^*

$$z^* = \frac{r_{sp} - 0}{1/\sqrt{n-1}} = r_{sp} \sqrt{n-1}$$

若 $|z^*| \geq z_{\frac{\alpha}{2}}$, 则拒绝 H_0 。

例题 15.8 学生能力测验分成数学和语文两部分。抽样 10 个学生的能力测验分数如表 15-7 所示。

表 15-7 学生的能力测验分数

| 学生 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 数学分数 | 425 | 358 | 515 | 672 | 378 | 397 | 715 | 638 | 478 | 350 |
| 语文分数 | 535 | 375 | 500 | 550 | 414 | 435 | 750 | 515 | 482 | 410 |

计算 Spearman 等级相关系数, 如表 15-8 所示。检验学生能力测验检验数学和语文两部分分数有无相关性, 如果检验的显著性水平是 0.05, 问检验的结果如何?

表 15-8 Spearman 等级相关系数

| 学生 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|--------------|----|---|---|---|---|---|----|---|---|----|
| 数学分数排序 r_i | 5 | 2 | 7 | 9 | 3 | 4 | 10 | 8 | 6 | 1 |
| 语文分数排序 s_i | 8 | 1 | 6 | 9 | 3 | 4 | 10 | 7 | 5 | 2 |
| d_i | -3 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | -1 |

解答: 以 r_i 当作变量 x_i , 以 s_i 当作变量 y_i , 则

$$S_{xy} = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} = 378 - 10(5.5)(5.5) = 75.5$$

$$S_{xx} = \sum_{i=1}^n x_i^2 - n \bar{x}^2 = 385 - 10(5.5)^2 = 82.5$$

$$S_{yy} = \sum_{i=1}^n y_i^2 - n \bar{y}^2 = 385 - 10(5.5)^2 = 82.5$$

$$r_{sp} = \frac{S_{rs}}{\sqrt{S_{rr} S_{ss}}} = \frac{75.5}{\sqrt{(82.5)(82.5)}} = 0.91515$$

$$r_{sp} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6(14)}{10(100 - 1)} = 0.91515$$

x_i 数据没有相等, 而且 y_i 数据也没有相等, 所以两个 r_{sp} 计算结果相同。

检验: $\begin{cases} H_0: \rho_s = 0, & \text{变量 } X \text{ 与变量 } Y \text{ 无线性相关性} \\ H_A: \rho_s \neq 0, & \text{变量 } X \text{ 与变量 } Y \text{ 有线性相关性} \end{cases}$

$$z^* = r_{sp} \sqrt{n-1} = (0.91515) \sqrt{10-1} = 2.75$$

因为 $|z^*| = 2.75 \geq z_{0.025} = 1.96$, 所以拒绝 H_0 。 p 值 = 0.0058。

查表 A-12: $r_\alpha = 0.648, r_{sp} > r_\alpha$, 所以拒绝 H_0 。

例题 15.9 (见网络资源)

15.9 中文统计应用

15.9.1 符号检验——单组样本（例题 15.1）

执行“符号检验——单组样本”的操作示意图和结果如图 15-4 所示。

| | | | | |
|----|----|----|----|----|
| | A | B | C | D |
| 1 | 22 | 23 | 25 | 25 |
| 2 | 30 | 32 | 25 | 25 |
| 3 | 31 | 20 | 25 | 25 |
| 4 | 40 | 40 | 25 | 25 |
| 5 | 37 | 28 | 25 | 25 |
| 6 | 25 | 26 | 25 | 25 |
| 7 | 29 | 33 | 25 | 25 |
| 8 | 14 | 25 | 25 | 25 |
| 9 | 30 | 34 | 25 | 25 |
| 10 | 17 | 21 | 25 | 25 |

符号检验

输入

样本1区域: A1:B10

样本2区域: C1:D10

使用方法: ☐ 正态分布 ☒ 二项分布

显著性水平: 0.05

☐ 标志

输出选项

| | | | | |
|----|-----------|---|---|----------|
| | A | B | C | D |
| 1 | 符号检验 正态分布 | | | |
| 2 | | | | |
| 3 | | | | |
| 4 | 大于零的样本个数 | | | 12 |
| 5 | 小于零的样本个数 | | | 6 |
| 6 | 等于零的样本个数 | | | 2 |
| 7 | 显著性水平 | | | 0.05 |
| 8 | z值 | | | 1.178511 |
| 9 | 单尾p值 | | | 0.1193 |
| 10 | 单尾z临界值 | | | 1.6449 |
| 11 | 双尾p值 | | | 0.2386 |
| 12 | 双尾z临界值 | | | 1.96 |

| | | | | |
|---|-----------|---|---|--------|
| | A | B | C | D |
| 1 | 符号检验 二项分配 | | | |
| 2 | | | | |
| 3 | | | | |
| 4 | 大于零的样本个数 | | | 12 |
| 5 | 小于零的样本个数 | | | 6 |
| 6 | 等于零的样本个数 | | | 2 |
| 7 | 显著性水平 | | | 0.05 |
| 8 | 单尾p值 | | | 0.1189 |
| 9 | 双尾p值 | | | 0.2378 |

图 15-4 执行“符号检验——单组样本”的操作示意图和结果

15.9.2 符号检验——两组配对样本（例题 15.2）

执行“符号检验——两组配对样本”的操作示意图和结果如图 15-5 所示。

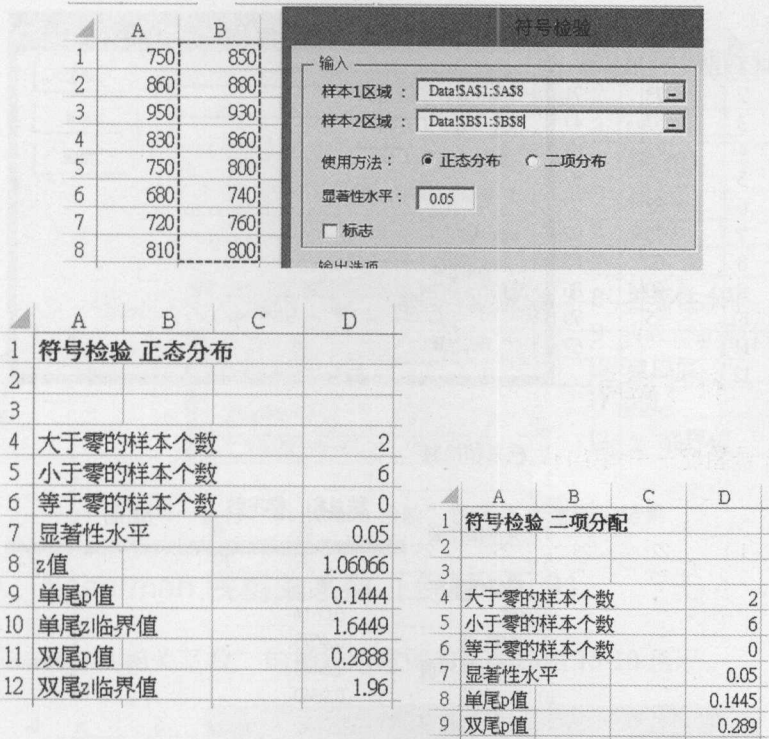


图 15-5 执行“符号检验——两组配对样本”的操作示意图和结果

15.9.3 游程检验（例题 15.4）

执行“游程检验”的操作示意图和结果如图 15-6 所示。

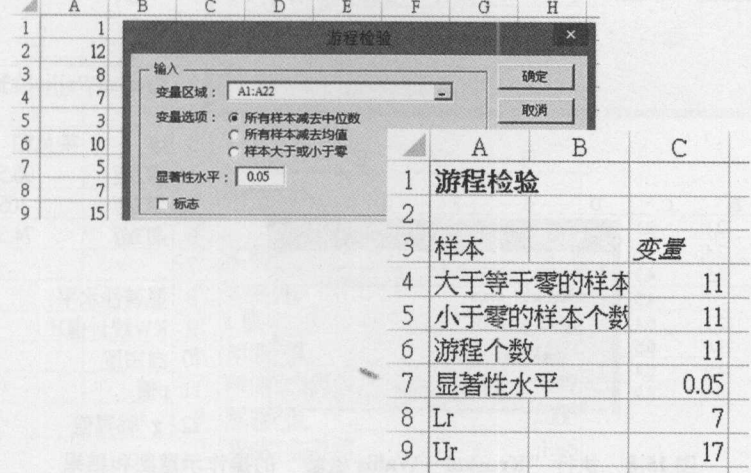


图 15-6 执行“游程检验”的操作示意图和结果

15.9.4 Mann – Whitney 秩总和检验 (例题 15.5)

执行“Mann – Whitney 秩总和检验”的操作示意图和结果如图 15-7 所示。

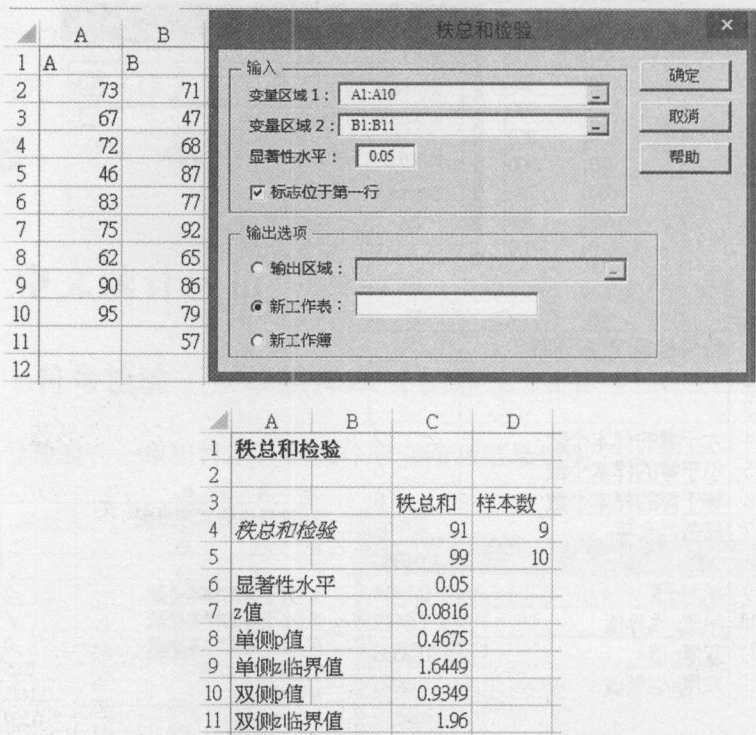


图 15-7 执行“Mann – Whitney 秩总和检验”的操作示意图和结果

15.9.5 Kruskal – Wallis 检验 (例题 15.6)

执行“Kruskal – Wallis 检验”的操作示意图和结果如图 15-8 所示。

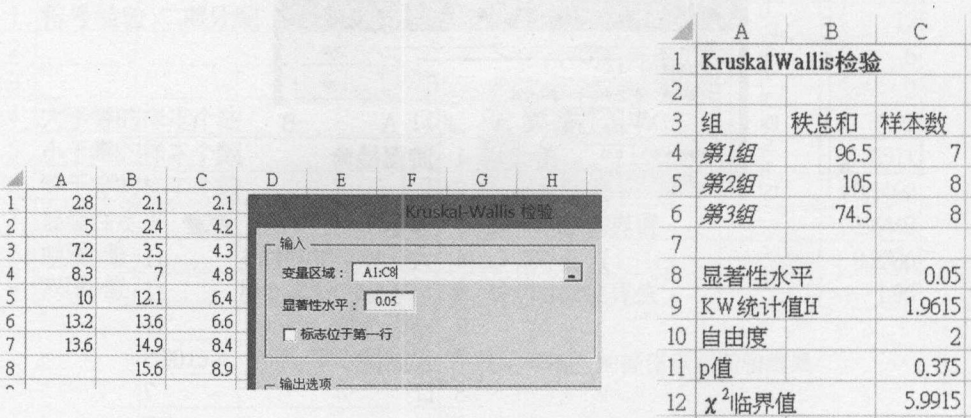


图 15-8 执行“Kruskal – Wallis 检验”的操作示意图和结果

15.9.6 Friedman 检验 (例题 15.7)

执行“Friedman 检验”的操作示意图和结果如图 15-9 所示。

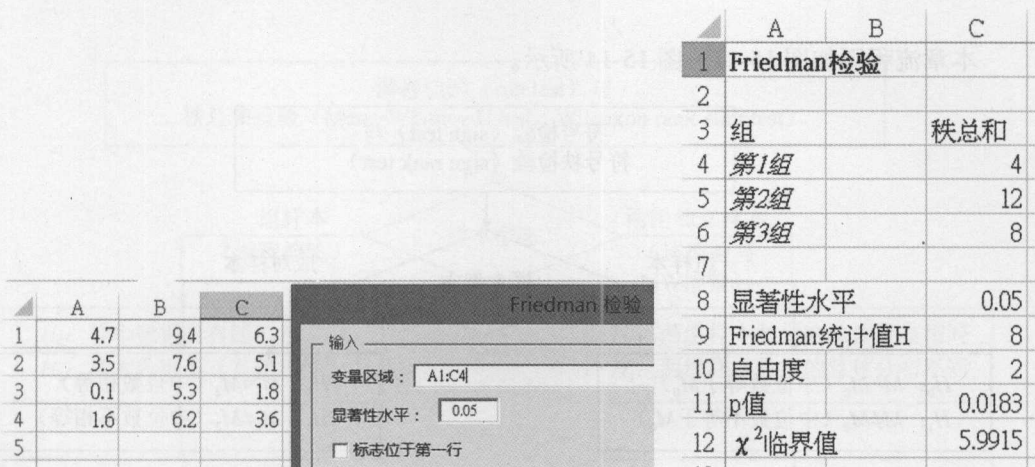


图 15-9 执行“Friedman 检验”的操作示意图和结果

15.9.7 Spearman 秩相关系数 (例题 15.8)

执行“Spearman 秩相关系数”的操作示意图和结果如图 15-10 所示。

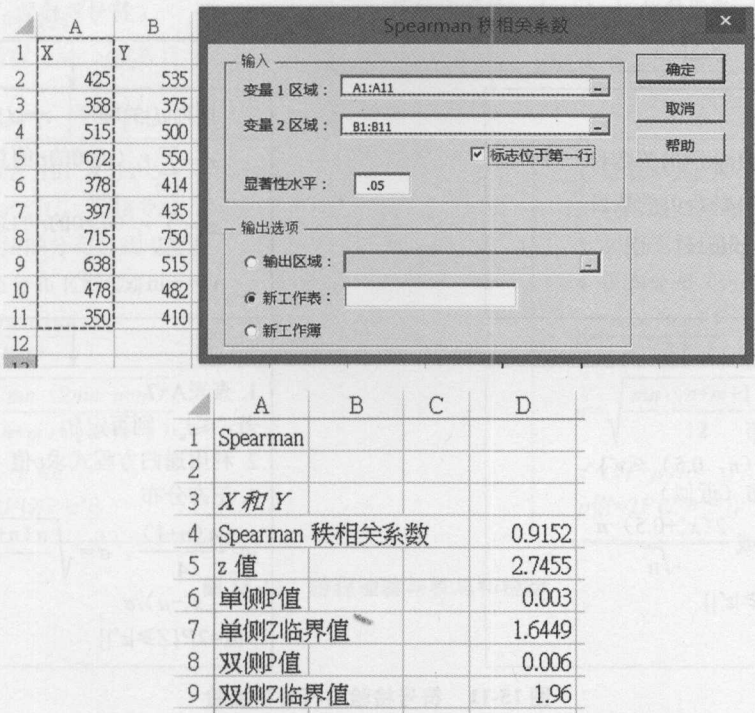


图 15-10 执行“Spearman 秩相关系数”的操作示意图和结果

15.10 本章流程图

本章流程图如图 15-11 ~ 图 15-14 所示。

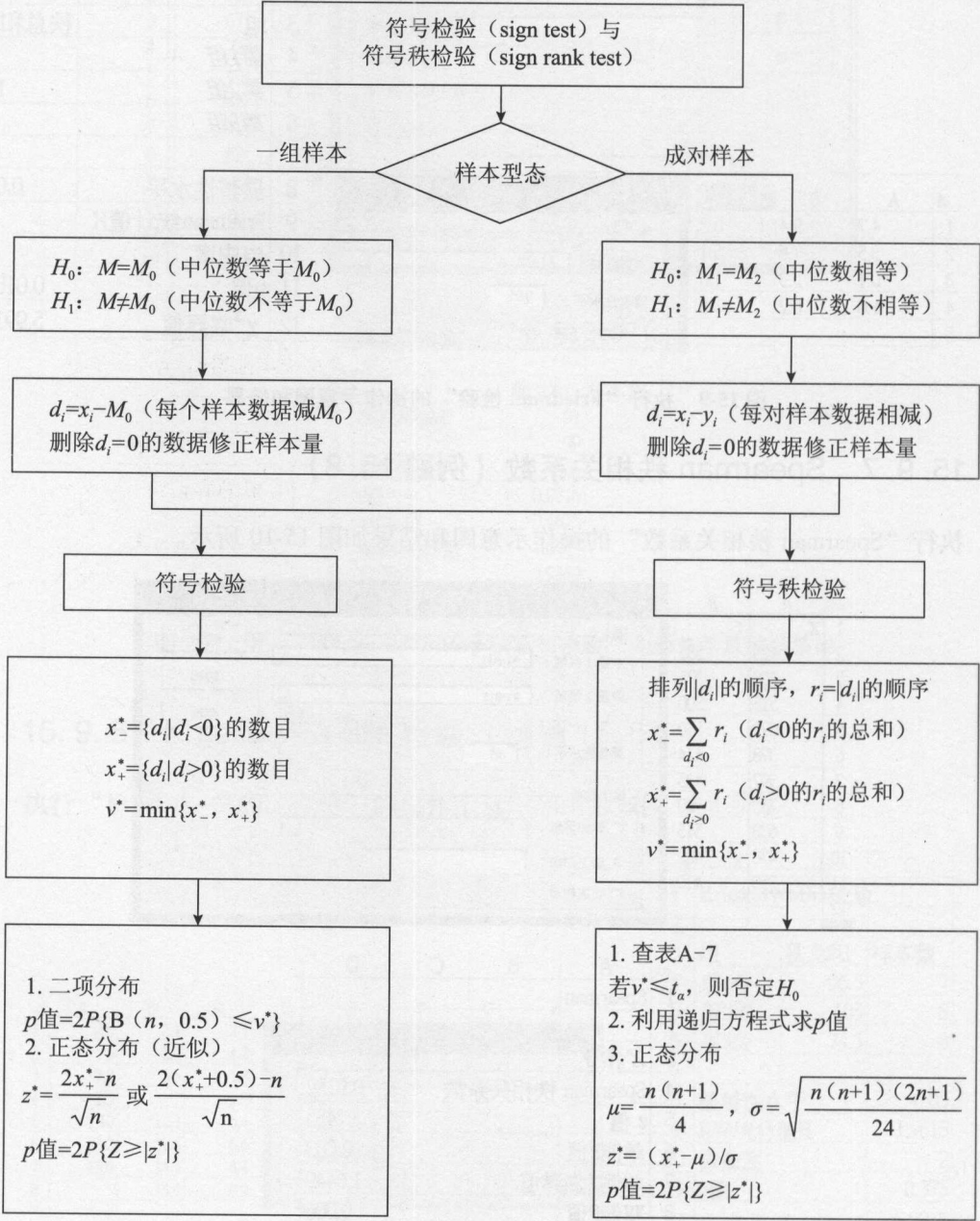
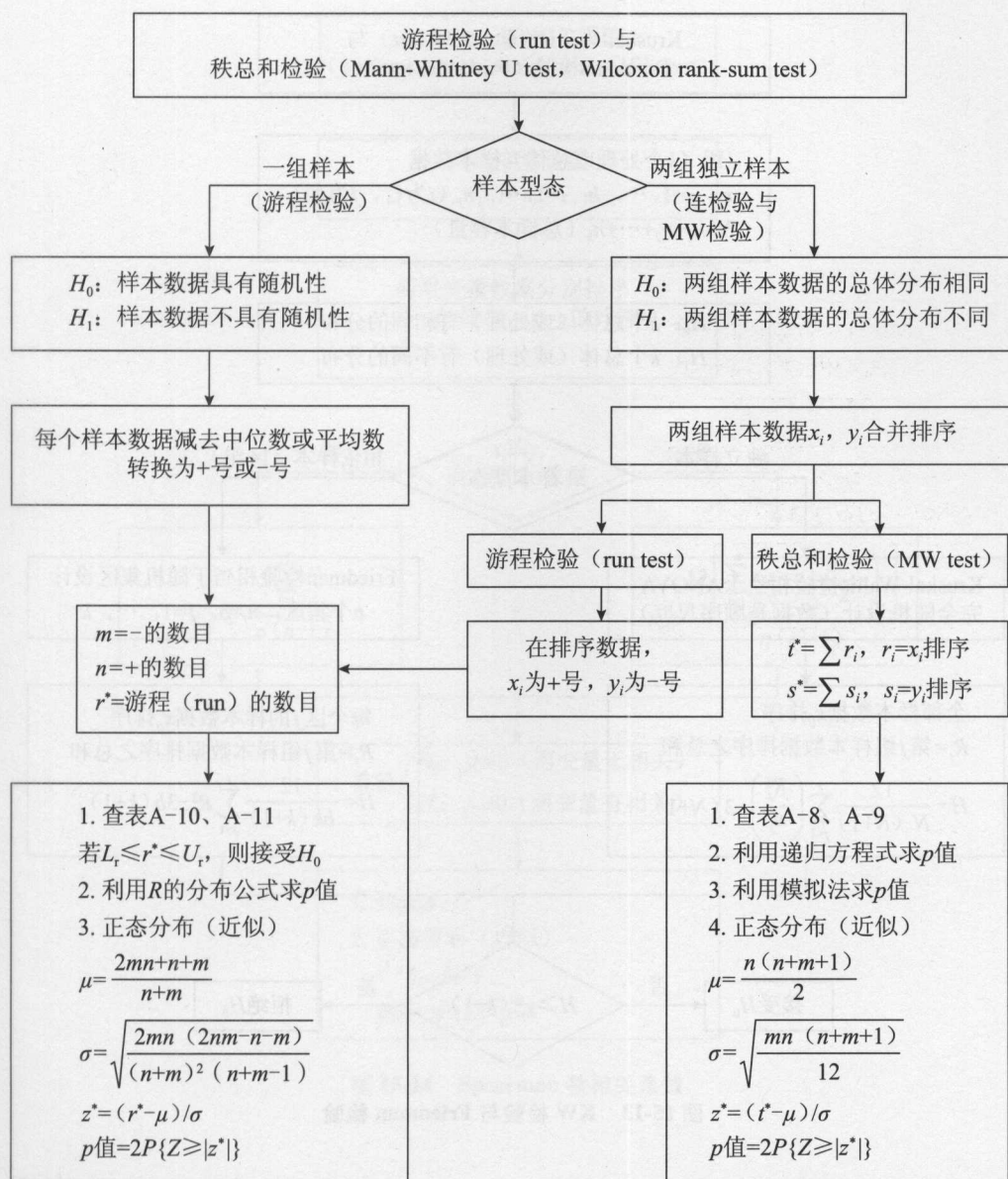


图 15-11 符号检验与符号秩检验



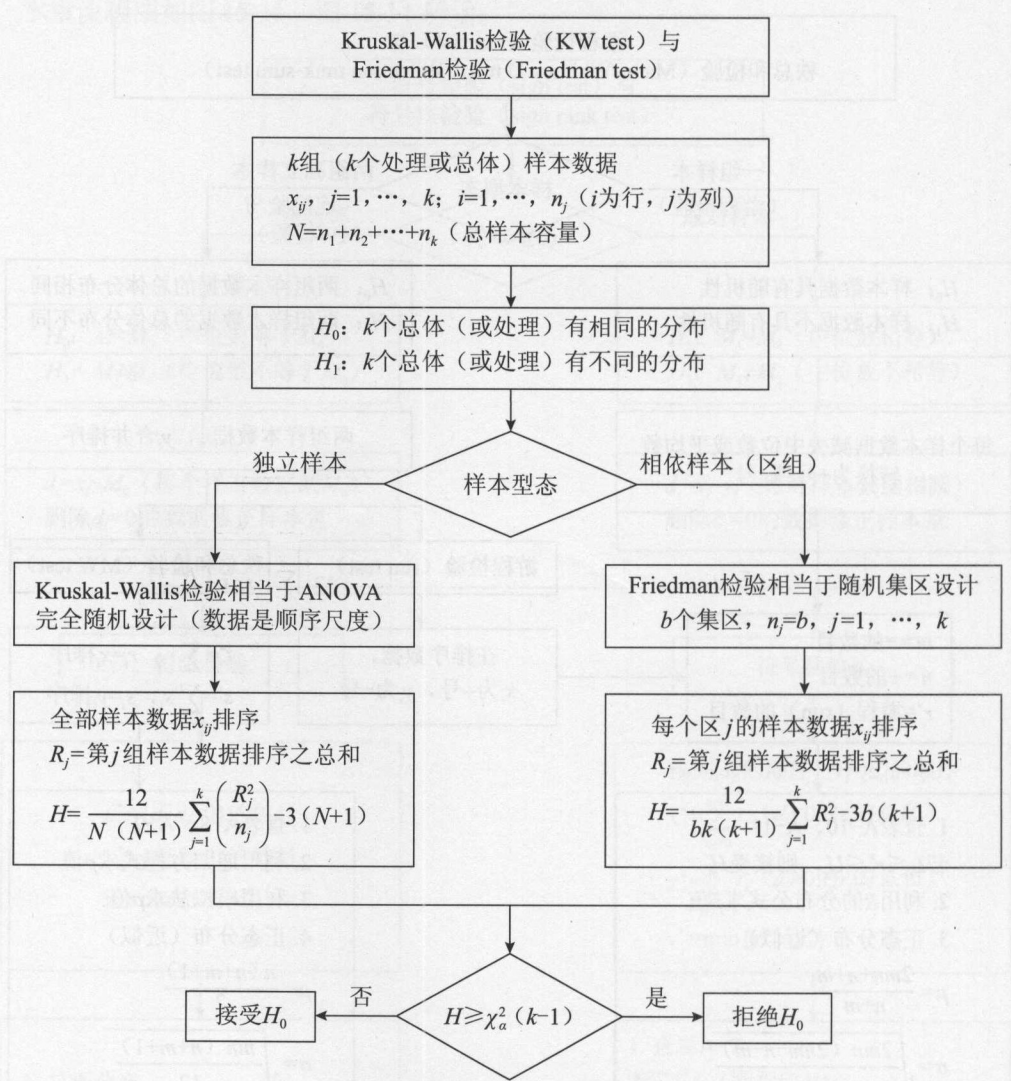


图 15-13 KW 检验与 Friedman 检验

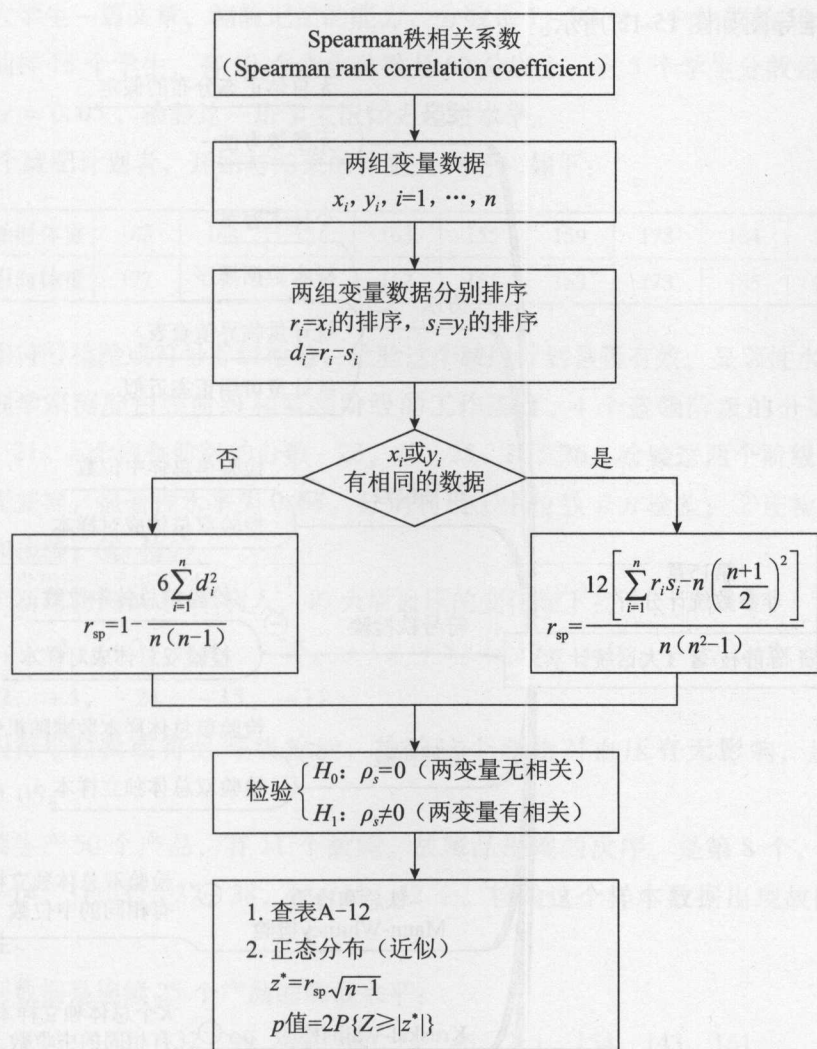


图 15-14 Spearman 秩相关系数

15.11 本章思维导图

本章思维导图如图 15-15 所示。

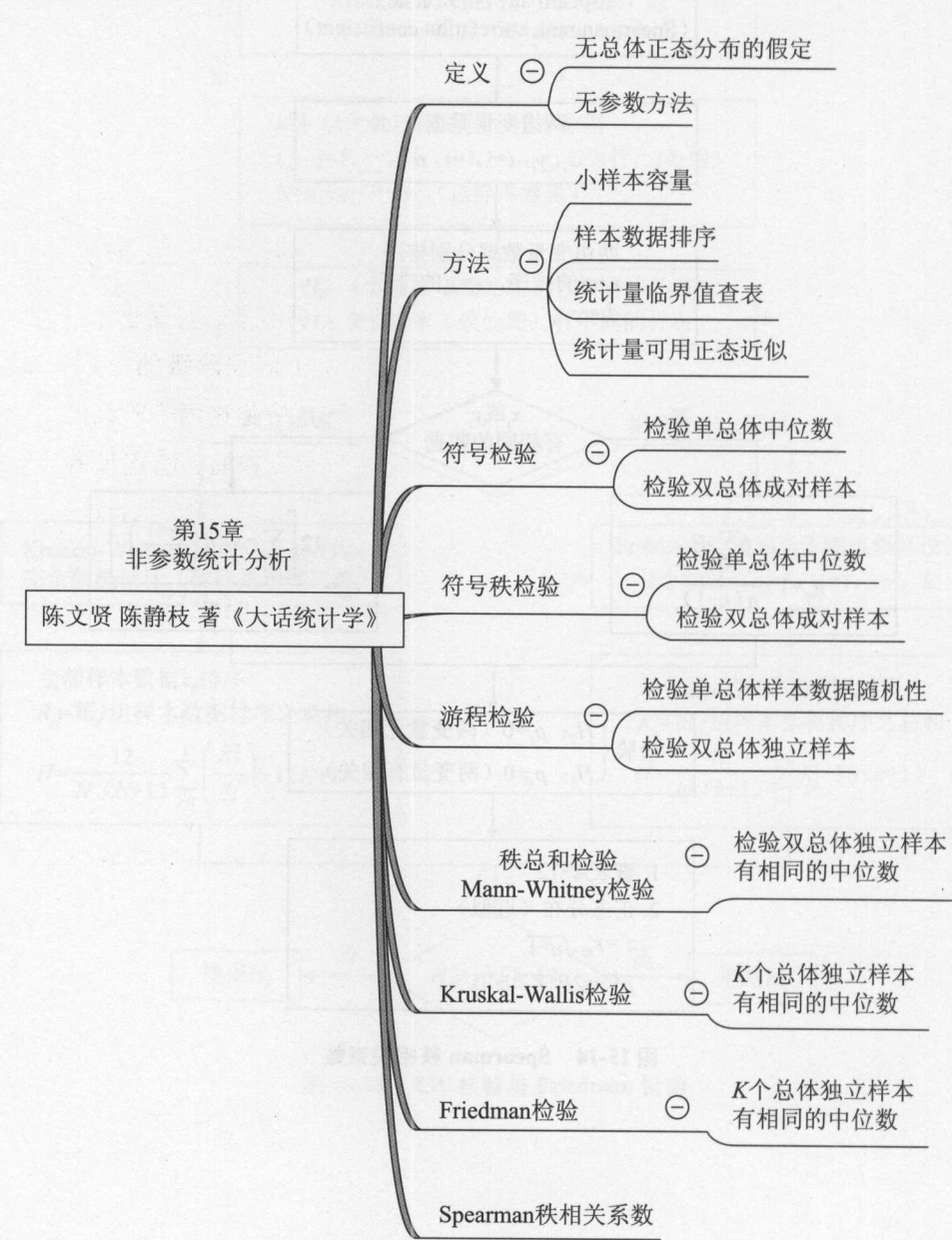


图 15-15 第 15 章思维导图

习 题

1. 给大学生一篇文章，测验记忆的能力，分数是 1 ~ 99 分，中位数是 50 分。现在一班抽样 15 个学生，有 10 个学生分数是 50 分以上，有 5 个学生分数是 50 分以下。以 $\alpha = 0.05$ ，检验这一班学生记忆力超过水平。

2. 10 个减肥计划者，开始与结束的体重（英磅）如下：

| | | | | | | | | | | |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 开始时体重 | 183 | 144 | 151 | 163 | 155 | 159 | 178 | 184 | 142 | 137 |
| 结束时体重 | 177 | 145 | 145 | 162 | 151 | 163 | 173 | 185 | 139 | 138 |

利用符号检验或符号等级检验，检验这个减肥计划是否有效，显著性水平为 0.05。

3. 心理学家测验白领阶级和蓝领阶级的工作态度，4 个蓝领阶级的分数：23，18，22，21。5 个白领阶级的分数：23，28，25，24，26。检验这两个阶级的工作态度有无差异，显著性水平为 0.05。分别利用①中位数卡方检验；②连检验；③等级总和检验；④t 检验。

4. 一个新药物测试 18 个病人，40 天后血压的变化如下：

-5， -1， +2， +8， -25， +1， +5， -12， -16， -9， -8， -18， -5，
-22， +4， -21， -15， -11

利用符号检验或符号等级检验，检验这个药物对血压有无影响，显著性水平为 0.05。

5. 连续生产 50 个产品，有 11 个故障。故障品出现的次序，是第 8 个，以及第 12，13，14，31，32，37，38，40，41，42 个。检验这个样本数据出现故障品具有随机性。

6. 下列数据是连续 25 个产品的质量水平：

100， 110， 122， 132， 99， 96， 88， 75， 45， 211， 154， 143， 161

142， 99， 111， 105， 133， 142， 150， 153， 121， 126， 117， 155

(1) 检验这个样本数据具有随机性。

(2) 检验这个样本数据服从正态分布，平均数为 124，标准差为 33。

其他习题请下载。



第 16 章

结 语

佛祖慧眼观看，见那猴王风车子一般相似，不住只管前进。大圣行时，忽见有五根肉红柱子，撑着一股青气，他道：“此间乃尽头路了，这番回去，如来作证，灵霄宫定是我坐也。”又思量说：“且住 A 等我留下些记号，方好与如来说话。”拔下一根毫毛，吹口仙气，叫“变 A”变作一管浓墨双毫笔，在那中间柱子上写一行大字云：“齐天大圣，到此一游。”

——吴承恩《西游记》

若你呼唤山，而山不来，你便应向它走去。

——《古兰经》

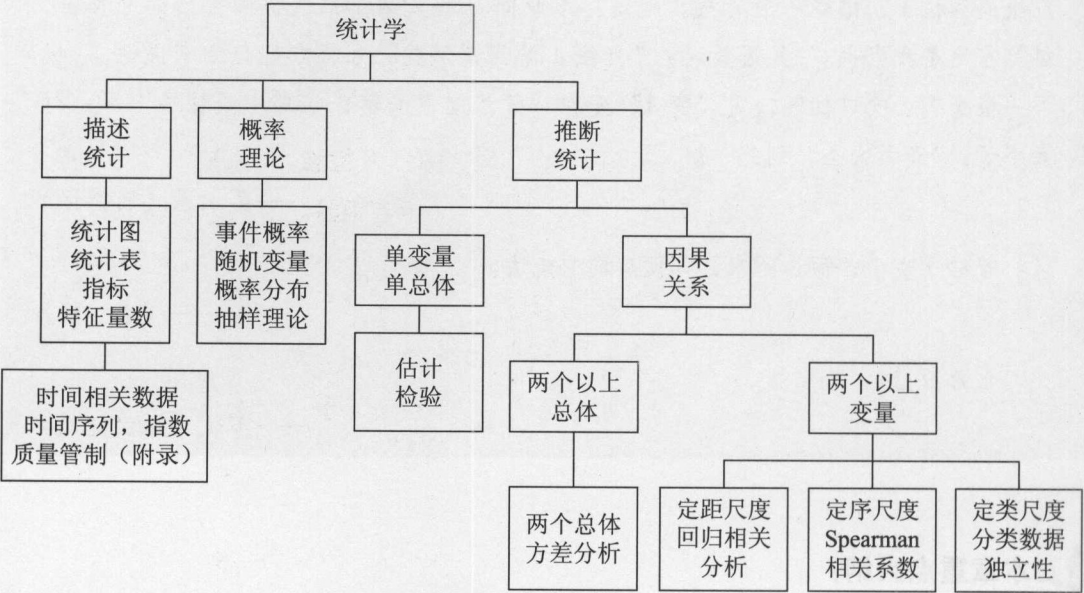
水落石出。

——苏轼《后赤壁赋》



本章重点大纲：

- 16.1 统计问题分类
- 16.2 误差名词说明
- 16.3 参数与统计量
- 16.4 统计概念复习



本章概念图

16.1 统计问题分类

1. 一个变量

(1.1) 一个定距变量：一个总体均值与方差的推断

(1.1.1) 一个总体均值的推断—— z 检验, t 检验。

(1.1.2) 一个总体方差的推断—— χ^2 分布检验。

(1.2) 一个定序变量。

(1.2.1) 一个总体中位数的推断——符号检验, 符号秩检验。

(1.2.2) 一个总体分布的推断——Kolmogorov, Lilliefors 检验。

(1.3) 一个定类变量。

(1.3.1) 一个总体比例的推断—— z 检验, 二项检验, F 检验。

(1.3.2) 一个总体多项比例或拟合优度的推断——卡方检验。

(1.3.3) 一个总体样本的随机性检验——游程检验。

2. 两个变量

(2.1) 两个定类变量。

(2.1.1) 一个定类变量 (标志分两个总体), 一个定类变量:

两个独立总体比例差的推断—— z 检验, 超几何检验。

(2.1.2) 两个配对总体比例的检验——McNemar 检验。

(2.1.3) 两个定类变量, 列联表: 分类数据独立性检验——卡方检验。

(2.2) 一个定类变量 (标志分类总体), 一个定序变量。

(2.2.1) 两个总体独立样本 (定序数据) ——游程检验, Mann - Whitney 检验, 中位数卡方检验。

(2.2.2) 两个总体配对样本 (定序数据) ——中位数检验, 符号检验, 符号秩检验

(2.2.3) 多个总体独立样本 (定序数据) ——KW 检验,

(2.2.4) 多个总体配对样本 (定序数据) ——Friedman 检验

(2.3) 一个定类变量 (标志分类总体), 一个定距变量。

(2.3.1) 两个总体独立样本均值差的推断—— z 检验, t 检验。

(2.3.2) 两个总体配对样本均值差的推断—— t 检验。

(2.3.3) 多个总体独立样本均值的推断——单因素方差分析。

(2.3.4) 多个总体区组样本均值的推断——双因素 (不重复) 方差分析。

- (2.3.5) 两个独立总体方差比的推断——F 检验。
- (2.3.6) 多个独立总体方差相等的检验——Bartlett 检验。
- (2.3.7) 因变数是定类变量，自变数是定距变量——逻辑回归。
- (2.4) 两个定序变量。
 - (2.4.1) Spearman 秩相关分析。
 - (2.4.2) 非参数回归。
- (2.5) 两个定距变量。
 - (2.5.1) 一元回归和相关分析。
 - (2.5.2) 时间序列分析。
- 3. 三个以上变量
 - (3.1) 多因素方差分析。
 - (3.2) 多元回归分析。
 - (3.3) 指数（变量：时间、商品、数量、价格）。
 - (3.4) 多变量分析。

以上统计问题分类，请对照图 16-1 和表 16-1。

表 16-1 统计问题分类

| | 一个变量 | | | 两个变量 | | | | |
|------|--------------------|----------|----------------|--------------------------|----------|--------------|-----------|-------------|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| 变量型态 | 定类尺度 | 定序尺度 | 定距尺度 | 两个定类 | 定类 & 定序 | 定类 & 定距 | 两个定序 | 两个定距 |
| 统计图表 | 直方图
频数表 | 排序数列 | 直方图
箱线图 | 列联表
条形图 | 多组
数列 | 多组
箱线图 | 时间
数列 | 散点图 |
| 统计量 | 百分比
(比例) | 中位数 | 均值
方差 | 比例差
$p_1 - p_2$ | 中位
数差 | 多组
平均数 | 秩相关
系数 | 共变量
相关系数 |
| 样本空间 | 间断型
{0, 1} | 离散型 | 连续型 | 联合间断 | 联合
间断 | 平面
坐标 | 联合
间断 | 联合间断
或连续 |
| 概率分布 | 贝努里
多项 | 任意
离散 | 正态
非正态 | 列联表 | 列联表 | 图 12-2 | 列联表 | 双正态 |
| 估计 | 比例值
二项 p | | 平均数
方差 | 比例差
$p_1 - p_2$ | | 均值差
方差比 | 秩相关
系数 | 回归估计 |
| 检验 | 二项 p
多项 p_i | 中位数 | 平均
方差
正态 | $p_1 - p_2$
卡方
独立性 | 中位数
差 | 均值差
ANOVA | 秩相关
系数 | 回归检验 |

| | | | | |
|-------|---------|---------------------------------|------------------------------|---------|
| 参数统计 | 单总体样本 | 检验均值 μ | z 检验, t 检验 | (1.1.1) |
| | | 检验比例 π | z , 二项 F 分布, χ^2 检验 | (1.3.1) |
| | | 检验方差 σ^2 | χ^2 分布 | (1.1.2) |
| | 双总体独立样本 | 检验均值差 $\mu_1 - \mu_2$ | z 检验, t 检验 | (2.3.1) |
| | | 检验比例差 $\pi_1 - \pi_2$ | z 检验 | (2.1.1) |
| | | 检验方差比 σ_1^2 / σ_2^2 | F 检验 | (2.3.5) |
| | 双总体配对样本 | 检验均值差 $\mu_1 - \mu_2$ | t 检验 | (2.3.2) |
| | 多总体独立样本 | 检验均值全相等 | 单因素方差分析 | (2.3.3) |
| | | 检验方差全相等 | Bartlett 检验 | (2.3.6) |
| | 多总体配对样本 | 检验均值全相等 | 二因素方差分析 无重复 | (2.3.4) |
| | 两组以上变量 | 回归与相关分析 | 一元回归 | (2.5.1) |
| | | 多变量分析 (3.4) | 多元回归 (3.2) | |
| | 单总体样本 | 卡方检验 | 多项分布 | (1.3.2) |
| | | | 拟合优度 | (1.3.2) |
| | | | 列联表独立性 | (2.1.3) |
| | | 检验中位数 | 符号检验 | (1.2.1) |
| | | | 符号秩检验 | |
| | | 检验随机性 | 游程检验 | (1.3.3) |
| | | 检验总体分布 | Kolmogorov 检验 | (1.2.2) |
| | | | Lilliefors 检验 | |
| | | 置信区间估计 | 自力法 | |
| | | | 重迭法 | |
| 非参数统计 | 双总体独立样本 | 检验比例相等 | 超几何分布 | (2.1.1) |
| | | 检验中位数相等 | 中位数卡方或游程检验 | (2.2.1) |
| | | 检验相同分布 | Mann-Whitney 检验 | (2.2.1) |
| | 双总体配对样本 | 检验中位数相等 | 符号或符号秩检验 | (2.2.2) |
| | | 检验比例相等 | McNemar 检验 | (2.1.2) |
| | 多总体独立样本 | 检验相同分布 | Kruskal-Wallis 检验 | (2.2.3) |
| | 多总体配对样本 | 检验相同分布 | Friedman 检验 | (2.2.4) |
| | 两组以上变量 | 相关分析 | Spearman 秩相关系数 | (2.4.1) |
| | | 回归分析 | Brown-Mood 法或 Theil 法 | (2.4.2) |
| | | 时间序列, (2.5.2) | 指数 (3.3) | |

图 16-1 统计问题分类图

图 16-2 为检验方法的关联, 小圆圈表示大圆圈的一个特例。

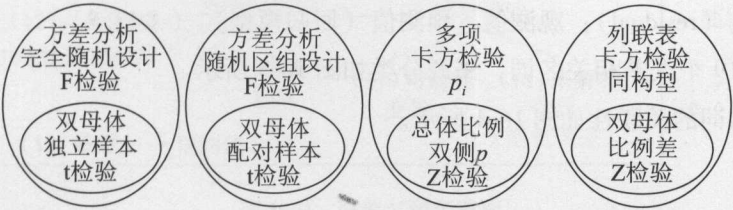


图 16-2 检验方法的关联图

16.2 误差名词说明

有关“误差”的名词及其出现的章节整理如下。

- (1) 非抽样误差 (non-sampling error): 因为取样和记录数据的误差。(§ 1.4)
 - (2) 抽样误差 (sampling error): 因为随机样本的误差。(§ 1.4)
 - (3) 差异 (difference): 两个数值 (观测值或统计值) 之差。
 - (4) 离差 (deviation): 观测值 - 平均数, 差异包括离差。(§ 2.6, § 12.3)
 - (5) 变异 (variation): 离差的平方和 (因为离差有正负)。(§ 12.2)
 - (6) 方差 (variance): 离差平方和的平均 (除以自由度) $= \sigma^2$ 。(§ 2.7.3, § 6.4)
 - (7) 离散程度 (variability): 数据变化分散的程度, 或称离差趋势。(§ 2.6)
 - (8) 离散变量 (discrete variable): 间断型数据的变量, 与误差无关。(§ 1.8)
 - (9) 置信水平 (confidence level) $(1 - \alpha)$: “估计”结果是“正确” (置信区间包含参数值) 的百分比。
 - (10) 显著性水平 (significant level) (α) : “检验”结果 (拒绝原假设) 是“错误”的概率之上限。
 - (11) 估计误差: 估计值 - 真实值 (参数) = 样本指标与总体指标之差 (§ 1.4), 误差变量 (§ 12.2)。
 - (12) 检验误差 (testing error): 假设检验的错误 (第一类, 第二类错误)。(§ 10.1)
 - (13) 偏误 (bias): 估计量期望值 - 参数 $= E(\text{估计量}) - \text{参数}$ 。(§ 9.2)
 - (14) 均方误 (mean square error): $E(\text{估计量} - \text{参数})^2 = \text{MSE}(W)$ 。(§ 2.6)
 - (15) 标准差 (standard deviation): 方差的平方根 $= \sigma$ 。(§ 2.6.2, § 6.4)
 - (16) 标准误差 (standard error): 统计量 (如 X) 的标准差。(§ 9.9)
 - (17) 可解释变异: 总体分群或变量因果关系模型可以解释的变异。(§ 13.5)
 - (18) 未解释变异: 方差分析或回归模型误差项的均方 $= \text{MSE}$ 。
 - (19) 残差 (residual): 观测值 - 预测值 (回归模型)。(§ 13.8)
- 将前述的 19 个误差相关名词, 来源分类如图 16-3 所示。
- “误差”名词的关联性如图 16-4 所示。

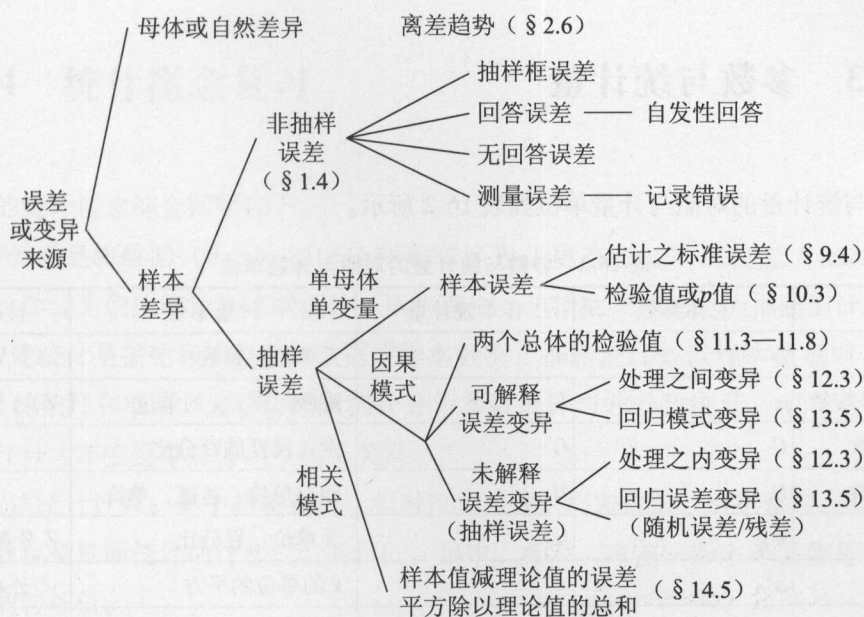


图 16-3 误差来源分类

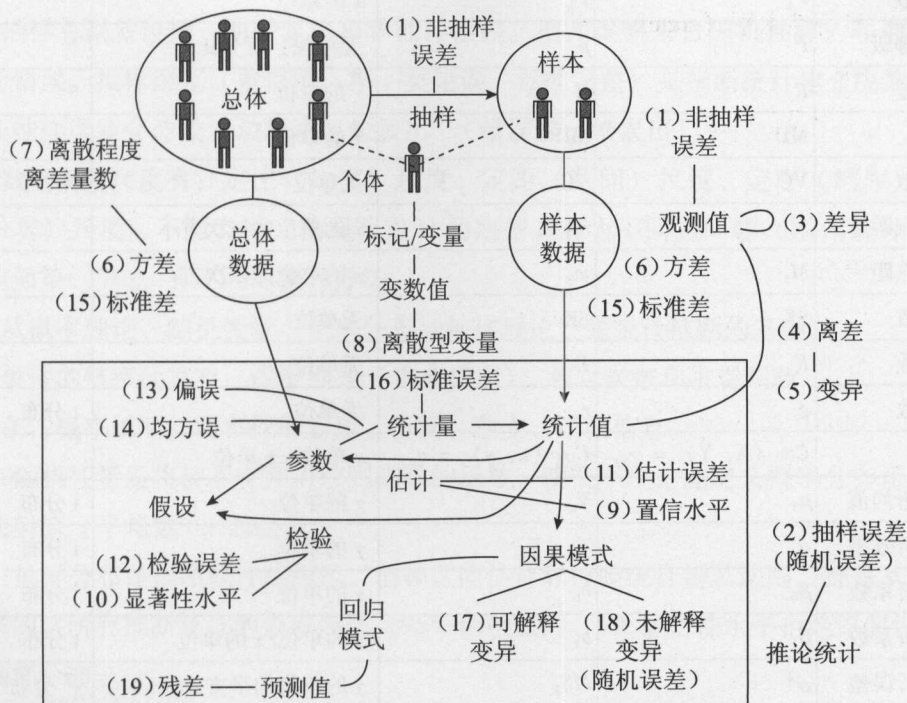


图 16-4 误差名词示意图

16.3 参数与统计量

参数与统计量的对照与计量单位如表 16-2 所示。

表 16-2 参数与统计量的对照与计量单位

| | 总体参数 | 样本统计量 | 计量单位 | 推断统计 |
|----------|---------------------------|-----------------------|----------------------|-------------|
| 容量 | N | n | 无单位：整数 | |
| (算术) 平均数 | μ | \bar{x} | 定距尺度： x 的单位 | Z, t 分布 |
| 几何平均数 | G | G | 定比尺度的百分比 | |
| 调和平均数 | H | H | 相对单位：速度，单价 | |
| 比例 | π | p | 无单位：百分比 | Z 分布 |
| 方差 | σ^2 | s^2 | x 的单位的平方 | χ^2 分布 |
| 标准差 | σ | s | x 的单位 | |
| 中位数 | M_e | M_e | x 的单位 | 非参数统计 |
| 百分位数 | P_k | P_k | x 的单位 | |
| 百分比等级 | P | p | 无单位：百分比 | |
| 全距极差 | R | R | x 的单位 | |
| 平均差 | MD | MD | x 的单位 | |
| 变异系数 | VC | VC | 无单位 | |
| 三阶原点距 | M_3' | m_3' | x 的单位的三次方 | |
| 四阶原点距 | M_4' | m_4' | x 的单位的四次方 | |
| 偏度系数 | SK | SK | 无单位 | |
| 峰度系数 | K | K | 无单位 | |
| 相关系数 | ρ | r | 无单位 | t 分布 |
| 协方差 | $Cov(X, Y) = \sigma_{XY}$ | $Cov(x, y) = q_{XY}$ | x 单位 $\times y$ 单位 | |
| 方差分析均值 | μ_i | \bar{y}_i | y 的单位 | t 分布 |
| 方差分析效果 | α_i | $\bar{y}_i - \bar{y}$ | y 的单位 | t 分布 |
| 回归分析系数 | β_0 | b_0 | y 的单位 | t 分布 |
| 回归分析系数 | β_1 | b_1 | y 的单位/ x 的单位 | t 分布 |
| 回归分析误差 | σ^2 | MS_E | y 的单位的平方 | χ^2 分布 |
| 分类数据分析 | π_i | p_i | 无单位 | χ^2 分布 |
| 指数 | | I_p, I_q | 无单位 | |

16.4 统计概念复习

本书的统计概念综合回顾如下。

1) 统计学是将数据 (data) 加以处理和转换为“更有意义”的信息 (information)。通常将统计学分为叙述统计和推断统计。概率理论 (包括抽样理论) 为推断统计的基础。

2) 叙述统计是描述和摘要总体数据或样本数据; 推断统计是以样本数据对总体特性 (特征值) 的估计和推断。通常将推断统计分为参数统计与非参数统计。在本书中, 我们将推断统计分为单变量单总体以及因果关系。

3) 叙述统计计算: 集中趋势量数、相对位置量数、离差量数、形态量数、相关量数。而这些量数也是推断统计估计和检验的目的。集中、相对、离差、形态等量数是随机变数概率分布的特征值。

4) 抽样即收集样本数据的方法有: 观察、调查和实验。实验有控制分组, 双总体推断 (匹配样本)、方差分析、回归分析等, 都可能有控制因子 (变量), 也就是实验设计。

5) 抽样总误差包括: 抽样误差和非抽样误差。抽样误差来自随机抽样, 非抽样误差来自人为错误。抽样误差 (置信度、第一类错误、标准误差) 是推断统计建立决策准则的根据。非抽样误差会造成“垃圾进、垃圾出” (错误进、错误出)。

6) 数据衡量尺度有: 定比 (比率) 尺度、定距 (区间) 尺度、定序 (顺序) 尺度、定类 (分类) 尺度。不同类型 (衡量尺度) 的数据, 因为不同的问题 (应用统计学的问题请见前面第一节), 有不同的统计方法。

7) 从概率理论、随机变量、概率分布到抽样理论, 是统计推断的基础。

8) 事件的概率计算有: 逻辑推导 (古典概率)、相对次数和主观判断。

9) 条件概率说明: 两事件的互斥、负面、独立、正面信息。

10) 随机变量是将样本空间对应到实数的函数, 使概率理论能定义概率分布函数, 并且计算期望值 (平均数)、方差等。

11) 抽样分布是推断统计的根源。信赖区间估计和检验统计量及法则, 都是从抽样分布导出来的。所有推断统计的概率叙述 (95% 信赖区间、5% 显著性水平拒绝原假设), 都是根据抽样分布。

12) 中心极限定理: 当抽样的样本数相当大时 ($n > 30$), 抽样平均 (统计量) 会近似正态分布, 其平均数等于总体随机变数的平均数, 方差等于总体方差除以 n 。

13) 所有的统计推断是根据抽样分布。例如, 在多少的置信度 (概率), 置信区间包

括参数；在多少的显著性水平（概率），拒绝原假设。但是不能说：有多少的概率，参数会落在置信区间，因为参数是确定值，不会有概率。

14) 假设检验会有第一类错误和第二类错误。在固定样本数之下，第一类错误降低会增加第二类错误；第二类错误降低会增加第一类错误。

15) 假设检验是根据“原假设成立”，计算样本数据检验统计值，如果样本统计值和原假设期望出现值的差距太大，则拒绝原假设。如何判定差距是否太大，决定于显著性水平 α ，即型 I 误差的概率。

16) 降低总体方差或样本方差，可以增加统计检验、显著性水平的结论。

17) 如果样本数增加，收集数据的成本增加。但是更多（样本）数据，就是更多信息。

(1) 甲：信赖区间估计的区间范围会更小。

(2) 乙：假设检验（包括双总体平均数或比例值、方差分析、卡方检验、非参数检验）的第一类错误和第二类错误会同时降低。

(3) 丙：检验的显著性会增加。

(4) 丁：回归系数的估计会更准确。

(5) 戊：两变数的相关性会更显著。

18) 因此，固定显著性水平 α ，如果样本数相当大（注意检验值的分子有 \sqrt{n} ，会使检验值变大、 p 值变小），则可以拒绝原假设（参数值等于多少、两总体相等），即有显著性。请用中文统计 3.0 的“快速检验”验算。样本数相当大即可直接估计，假设检验则无多大意义。假设检验的目的是在有限的样本下，利用检验的结果，做出比较明智的决策。统计学的“路灯”，不只是“支持”，而且要在“黑夜”里（有限样本数目）“照明”。

19) 分析定距尺度的因（依赖）变量之样本信息，总变异（所有样本数据和总平均数之差的平方和）是固定的，而要从变异来源，检验自（独立）变量（方差分析的总体分类变量或回归分析的控制变量），是否可以影响因变量。

20) 观察或调查数据，可能导致冲突的解释，因为未定义的因素可能抵消自变量影响的效应。如果用实验设计可以增加实验或集区因子（例如成对样本），就是增加“变异来源”，可使检验自变量效应的显著性提高。

21) 方差分析是回归分析的特例，其自变量是虚变量（0—1 变量）。但是方差分析问题不会用回归分析求解，因为模型和计算会更复杂。

22) 利用有母数统计必须符合一些假定条件（正态、相同方差等），所以如果不确定符合这些假设条件，则要先检查这些条件是否成立（如回归的残差分析）。

23) 如果不符合有母数统计的假设条件，则可以利用非参数统计。非参数统计将原始

信息转为定序尺度。

24) 非参数统计利用定序尺度 (排列大小), 除了符号检验的统计量是二项分布, 其他统计量各有不同 (而且不是常用) 的概率分布 (表), 但是当样本数稍大 ($n > 10$ 或 15), 则可用正态分布或卡方分布来近似。

25) 非参数统计 (总体不是正态分布或没有参数) 的步骤。

(1) 将样本数据排列定序尺度随机变量。

(2) 定义定序尺度随机变量的统计量。

(3) 统计量的概率分布 (查表或近似正态)。

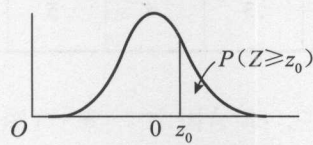
26) 因为 t 分布和 F 分布的关联性, 所以双总体平均数 t 检验是方差分析 F 检验的特例。还有 Z 分布和 χ^2 分布的关联性, 双总体比例值 Z 检验是多项间断卡方检验的特例。如图 16-1 所示。即

$$[t(n)]^2 = F(1, n), [t_{\alpha/2}(n)]^2 = F_{\alpha}(1, n) \quad (Z)^2 = \chi^2(1), (z_{\alpha/2})^2 = \chi_{\alpha}^2(1)$$

27) 有更多的钱 → 收集更多的数据 (没有非抽样误差、可以做实验设计、符合正态分布) → 得到更多的信息 (叙述统计、推断统计) → 抽样更小的标准误 → 可以制定更小的显著性水平 (型 I 误差) → 获得更好的决策 (显著性结论、表 10-2 的笑脸)。

附录一

正态分布概率表



$P(Z \geq z_0)$ 机率表

表 A-1 正态分布概率表

| z_0 小数第一位 | z_0 小数第二位 | | | | | | | | | |
|-------------|-------------|------|------|------|------|------|------|------|------|------|
| | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
| 0.0 | .500 | .496 | .495 | .488 | .484 | .480 | .476 | .472 | .468 | .464 |
| 0.1 | .460 | .456 | .452 | .448 | .444 | .440 | .436 | .433 | .429 | .425 |
| 0.2 | .421 | .417 | .413 | .409 | .405 | .401 | .397 | .394 | .390 | .386 |
| 0.3 | .382 | .378 | .374 | .371 | .367 | .363 | .359 | .356 | .352 | .348 |
| 0.4 | .345 | .341 | .337 | .334 | .330 | .326 | .323 | .319 | .316 | .312 |
| 0.5 | .309 | .305 | .302 | .298 | .295 | .291 | .288 | .284 | .281 | .278 |
| 0.6 | .274 | .271 | .268 | .264 | .261 | .258 | .255 | .251 | .248 | .245 |
| 0.7 | .242 | .239 | .236 | .233 | .230 | .227 | .224 | .221 | .218 | .215 |
| 0.8 | .212 | .209 | .206 | .203 | .200 | .198 | .195 | .192 | .189 | .187 |
| 0.9 | .184 | .181 | .179 | .176 | .174 | .171 | .169 | .166 | .164 | .161 |
| 1.0 | .159 | .156 | .154 | .152 | .149 | .147 | .145 | .142 | .140 | .138 |
| 1.1 | .136 | .133 | .131 | .129 | .127 | .125 | .123 | .121 | .119 | .117 |
| 1.2 | .115 | .113 | .111 | .109 | .107 | .106 | .104 | .102 | .100 | .099 |
| 1.3 | .097 | .095 | .093 | .092 | .090 | .089 | .087 | .085 | .084 | .082 |
| 1.4 | .081 | .079 | .078 | .076 | .075 | .074 | .072 | .071 | .069 | .068 |
| 1.5 | .067 | .066 | .064 | .063 | .062 | .061 | .059 | .058 | .057 | .056 |
| 1.6 | .055 | .054 | .053 | .052 | .051 | .049 | .048 | .047 | .046 | .046 |
| 1.7 | .045 | .044 | .043 | .042 | .041 | .040 | .039 | .038 | .038 | .037 |
| 1.8 | .036 | .035 | .034 | .034 | .033 | .032 | .031 | .031 | .030 | .029 |
| 1.9 | .029 | .028 | .027 | .027 | .026 | .026 | .025 | .024 | .024 | .023 |
| 2.0 | .023 | .022 | .022 | .021 | .021 | .020 | .020 | .019 | .019 | .018 |
| 2.1 | .018 | .017 | .017 | .017 | .016 | .016 | .015 | .015 | .015 | .014 |
| 2.2 | .014 | .014 | .013 | .013 | .013 | .012 | .012 | .012 | .011 | .011 |
| 2.3 | .011 | .010 | .010 | .010 | .010 | .009 | .009 | .009 | .009 | .008 |
| 2.4 | .008 | .008 | .008 | .008 | .007 | .007 | .007 | .007 | .007 | .006 |
| 2.5 | .006 | .006 | .006 | .006 | .006 | .005 | .005 | .005 | .005 | .005 |
| 2.6 | .005 | .005 | .004 | .004 | .004 | .004 | .004 | .004 | .004 | .004 |
| 2.7 | .003 | .003 | .003 | .003 | .003 | .003 | .003 | .003 | .003 | .003 |
| 2.8 | .003 | .002 | .002 | .002 | .002 | .002 | .002 | .002 | .002 | .002 |
| 2.9 | .002 | .002 | .002 | .002 | .002 | .002 | .002 | .001 | .001 | .001 |

续表

| z_0 整数 | $.1228 = .0228$ | | | | | $.135 = .00135$ | | | | |
|-------------|-----------------|--------|--------|--------|--------|-----------------|--------|--------|--------|--------|
| 2. | .1228 | .1179 | .1139 | .1107 | .10820 | .10621 | .10446 | .10347 | .10256 | .10187 |
| 3. | .12135 | .1168 | .11287 | .10983 | .10737 | .10533 | .10359 | .10268 | .10173 | .10101 |
| 4. | .11987 | .11527 | .11133 | .10829 | .10583 | .10379 | .10205 | .10114 | .10019 | .09947 |
| 5. | .11847 | .11387 | .10993 | .10689 | .10443 | .10239 | .10065 | .09974 | .09879 | .09807 |
| z_0 小数第一位 | .0 | .1 | .2 | .3 | .4 | .5 | .6 | .7 | .8 | .9 |

| | | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|------|---------|
| 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 0.1-0.9 |
| 0.01 | 0.04 | 0.09 | 0.14 | 0.19 | 0.24 | 0.29 | 0.34 | 0.39 | 0.44 | 0.5 |
| 0.02 | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.5 |
| 0.03 | 0.06 | 0.11 | 0.16 | 0.21 | 0.26 | 0.31 | 0.36 | 0.41 | 0.46 | 0.5 |
| 0.04 | 0.07 | 0.12 | 0.17 | 0.22 | 0.27 | 0.32 | 0.37 | 0.42 | 0.47 | 0.5 |
| 0.05 | 0.08 | 0.13 | 0.18 | 0.23 | 0.28 | 0.33 | 0.38 | 0.43 | 0.48 | 0.5 |
| 0.06 | 0.09 | 0.14 | 0.19 | 0.24 | 0.29 | 0.34 | 0.39 | 0.44 | 0.49 | 0.5 |
| 0.07 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 | 0.5 |
| 0.08 | 0.11 | 0.16 | 0.21 | 0.26 | 0.31 | 0.36 | 0.41 | 0.46 | 0.51 | 0.5 |
| 0.09 | 0.12 | 0.17 | 0.22 | 0.27 | 0.32 | 0.37 | 0.42 | 0.47 | 0.52 | 0.5 |
| 0.10 | 0.13 | 0.18 | 0.23 | 0.28 | 0.33 | 0.38 | 0.43 | 0.48 | 0.53 | 0.5 |
| 0.11 | 0.14 | 0.19 | 0.24 | 0.29 | 0.34 | 0.39 | 0.44 | 0.49 | 0.54 | 0.5 |
| 0.12 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 | 0.55 | 0.5 |
| 0.13 | 0.16 | 0.21 | 0.26 | 0.31 | 0.36 | 0.41 | 0.46 | 0.51 | 0.56 | 0.5 |
| 0.14 | 0.17 | 0.22 | 0.27 | 0.32 | 0.37 | 0.42 | 0.47 | 0.52 | 0.57 | 0.5 |
| 0.15 | 0.18 | 0.23 | 0.28 | 0.33 | 0.38 | 0.43 | 0.48 | 0.53 | 0.58 | 0.5 |
| 0.16 | 0.19 | 0.24 | 0.29 | 0.34 | 0.39 | 0.44 | 0.49 | 0.54 | 0.59 | 0.5 |
| 0.17 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 | 0.55 | 0.60 | 0.5 |
| 0.18 | 0.21 | 0.26 | 0.31 | 0.36 | 0.41 | 0.46 | 0.51 | 0.56 | 0.61 | 0.5 |
| 0.19 | 0.22 | 0.27 | 0.32 | 0.37 | 0.42 | 0.47 | 0.52 | 0.57 | 0.62 | 0.5 |
| 0.20 | 0.23 | 0.28 | 0.33 | 0.38 | 0.43 | 0.48 | 0.53 | 0.58 | 0.63 | 0.5 |
| 0.21 | 0.24 | 0.29 | 0.34 | 0.39 | 0.44 | 0.49 | 0.54 | 0.59 | 0.64 | 0.5 |
| 0.22 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 | 0.55 | 0.60 | 0.65 | 0.5 |
| 0.23 | 0.26 | 0.31 | 0.36 | 0.41 | 0.46 | 0.51 | 0.56 | 0.61 | 0.66 | 0.5 |
| 0.24 | 0.27 | 0.32 | 0.37 | 0.42 | 0.47 | 0.52 | 0.57 | 0.62 | 0.67 | 0.5 |
| 0.25 | 0.28 | 0.33 | 0.38 | 0.43 | 0.48 | 0.53 | 0.58 | 0.63 | 0.68 | 0.5 |
| 0.26 | 0.29 | 0.34 | 0.39 | 0.44 | 0.49 | 0.54 | 0.59 | 0.64 | 0.69 | 0.5 |
| 0.27 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 | 0.55 | 0.60 | 0.65 | 0.70 | 0.5 |
| 0.28 | 0.31 | 0.36 | 0.41 | 0.46 | 0.51 | 0.56 | 0.61 | 0.66 | 0.71 | 0.5 |
| 0.29 | 0.32 | 0.37 | 0.42 | 0.47 | 0.52 | 0.57 | 0.62 | 0.67 | 0.72 | 0.5 |
| 0.30 | 0.33 | 0.38 | 0.43 | 0.48 | 0.53 | 0.58 | 0.63 | 0.68 | 0.73 | 0.5 |
| 0.31 | 0.34 | 0.39 | 0.44 | 0.49 | 0.54 | 0.59 | 0.64 | 0.69 | 0.74 | 0.5 |
| 0.32 | 0.35 | 0.40 | 0.45 | 0.50 | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.5 |
| 0.33 | 0.36 | 0.41 | 0.46 | 0.51 | 0.56 | 0.61 | 0.66 | 0.71 | 0.76 | 0.5 |
| 0.34 | 0.37 | 0.42 | 0.47 | 0.52 | 0.57 | 0.62 | 0.67 | 0.72 | 0.77 | 0.5 |
| 0.35 | 0.38 | 0.43 | 0.48 | 0.53 | 0.58 | 0.63 | 0.68 | 0.73 | 0.78 | 0.5 |
| 0.36 | 0.39 | 0.44 | 0.49 | 0.54 | 0.59 | 0.64 | 0.69 | 0.74 | 0.79 | 0.5 |
| 0.37 | 0.40 | 0.45 | 0.50 | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.5 |
| 0.38 | 0.41 | 0.46 | 0.51 | 0.56 | 0.61 | 0.66 | 0.71 | 0.76 | 0.81 | 0.5 |
| 0.39 | 0.42 | 0.47 | 0.52 | 0.57 | 0.62 | 0.67 | 0.72 | 0.77 | 0.82 | 0.5 |
| 0.40 | 0.43 | 0.48 | 0.53 | 0.58 | 0.63 | 0.68 | 0.73 | 0.78 | 0.83 | 0.5 |
| 0.41 | 0.44 | 0.49 | 0.54 | 0.59 | 0.64 | 0.69 | 0.74 | 0.79 | 0.84 | 0.5 |
| 0.42 | 0.45 | 0.50 | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.5 |
| 0.43 | 0.46 | 0.51 | 0.56 | 0.61 | 0.66 | 0.71 | 0.76 | 0.81 | 0.86 | 0.5 |
| 0.44 | 0.47 | 0.52 | 0.57 | 0.62 | 0.67 | 0.72 | 0.77 | 0.82 | 0.87 | 0.5 |
| 0.45 | 0.48 | 0.53 | 0.58 | 0.63 | 0.68 | 0.73 | 0.78 | 0.83 | 0.88 | 0.5 |
| 0.46 | 0.49 | 0.54 | 0.59 | 0.64 | 0.69 | 0.74 | 0.79 | 0.84 | 0.89 | 0.5 |
| 0.47 | 0.50 | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.5 |
| 0.48 | 0.51 | 0.56 | 0.61 | 0.66 | 0.71 | 0.76 | 0.81 | 0.86 | 0.91 | 0.5 |
| 0.49 | 0.52 | 0.57 | 0.62 | 0.67 | 0.72 | 0.77 | 0.82 | 0.87 | 0.92 | 0.5 |
| 0.50 | 0.53 | 0.58 | 0.63 | 0.68 | 0.73 | 0.78 | 0.83 | 0.88 | 0.93 | 0.5 |
| 0.51 | 0.54 | 0.59 | 0.64 | 0.69 | 0.74 | 0.79 | 0.84 | 0.89 | 0.94 | 0.5 |
| 0.52 | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 | 0.5 |
| 0.53 | 0.56 | 0.61 | 0.66 | 0.71 | 0.76 | 0.81 | 0.86 | 0.91 | 0.96 | 0.5 |
| 0.54 | 0.57 | 0.62 | 0.67 | 0.72 | 0.77 | 0.82 | 0.87 | 0.92 | 0.97 | 0.5 |
| 0.55 | 0.58 | 0.63 | 0.68 | 0.73 | 0.78 | 0.83 | 0.88 | 0.93 | 0.98 | 0.5 |
| 0.56 | 0.59 | 0.64 | 0.69 | 0.74 | 0.79 | 0.84 | 0.89 | 0.94 | 0.99 | 0.5 |
| 0.57 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 | 1.00 | 0.5 |
| 0.58 | 0.61 | 0.66 | 0.71 | 0.76 | 0.81 | 0.86 | 0.91 | 0.96 | 1.01 | 0.5 |
| 0.59 | 0.62 | 0.67 | 0.72 | 0.77 | 0.82 | 0.87 | 0.92 | 0.97 | 1.02 | 0.5 |
| 0.60 | 0.63 | 0.68 | 0.73 | 0.78 | 0.83 | 0.88 | 0.93 | 0.98 | 1.03 | 0.5 |
| 0.61 | 0.64 | 0.69 | 0.74 | 0.79 | 0.84 | 0.89 | 0.94 | 0.99 | 1.04 | 0.5 |
| 0.62 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 | 1.00 | 1.05 | 0.5 |
| 0.63 | 0.66 | 0.71 | 0.76 | 0.81 | 0.86 | 0.91 | 0.96 | 1.01 | 1.06 | 0.5 |
| 0.64 | 0.67 | 0.72 | 0.77 | 0.82 | 0.87 | 0.92 | 0.97 | 1.02 | 1.07 | 0.5 |
| 0.65 | 0.68 | 0.73 | 0.78 | 0.83 | 0.88 | 0.93 | 0.98 | 1.03 | 1.08 | 0.5 |
| 0.66 | 0.69 | 0.74 | 0.79 | 0.84 | 0.89 | 0.94 | 0.99 | 1.04 | 1.09 | 0.5 |
| 0.67 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 | 1.00 | 1.05 | 1.10 | 0.5 |
| 0.68 | 0.71 | 0.76 | 0.81 | 0.86 | 0.91 | 0.96 | 1.01 | 1.06 | 1.11 | 0.5 |
| 0.69 | 0.72 | 0.77 | 0.82 | 0.87 | 0.92 | 0.97 | 1.02 | 1.07 | 1.12 | 0.5 |
| 0.70 | 0.73 | 0.78 | 0.83 | 0.88 | 0.93 | 0.98 | 1.03 | 1.08 | 1.13 | 0.5 |
| 0.71 | 0.74 | 0.79 | 0.84 | 0.89 | 0.94 | 0.99 | 1.04 | 1.09 | 1.14 | 0.5 |
| 0.72 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 | 1.00 | 1.05 | 1.10 | 1.15 | 0.5 |
| 0.73 | 0.76 | 0.81 | 0.86 | 0.91 | 0.96 | 1.01 | 1.06 | 1.11 | 1.16 | 0.5 |
| 0.74 | 0.77 | 0.82 | 0.87 | 0.92 | 0.97 | 1.02 | 1.07 | 1.12 | 1.17 | 0.5 |
| 0.75 | 0.78 | 0.83 | 0.88 | 0.93 | 0.98 | 1.03 | 1.08 | 1.13 | 1.18 | 0.5 |
| 0.76 | 0.79 | 0.84 | 0.89 | 0.94 | 0.99 | 1.04 | 1.09 | 1.14 | 1.19 | 0.5 |
| 0.77 | 0.80 | 0.85 | 0.90 | 0.95 | 1.00 | 1.05 | 1.10 | 1.15 | 1.20 | 0.5 |
| 0.78 | 0.81 | 0.86 | 0.91 | 0.96 | 1.01 | 1.06 | 1.11 | 1.16 | 1.21 | 0.5 |
| 0.79 | 0.82 | 0.87 | 0.92 | 0.97 | 1.02 | 1.07 | 1.12 | 1.17 | 1.22 | 0.5 |
| 0.80 | 0.83 | 0.88 | 0.93 | 0.98 | 1.03 | 1.08 | 1.13 | 1.18 | 1.23 | 0.5 |
| 0.81 | 0.84 | 0.89 | 0.94 | 0.99 | 1.04 | 1.09 | 1.14 | 1.19 | 1.24 | 0.5 |
| 0.82 | 0.85 | 0.90 | 0.95 | 1.00 | 1.05 | 1.10 | 1.15 | 1.20 | 1.25 | 0.5 |
| 0.83 | 0.86 | 0.91 | 0.96 | 1.01 | 1.06 | 1.11 | 1.16 | 1.21 | 1.26 | 0.5 |
| 0.84 | 0.87 | 0.92 | 0.97 | 1.02 | 1.07 | 1.12 | 1.17 | 1.22 | 1.27 | 0.5 |
| 0.85 | 0.88 | 0.93 | 0.98 | 1.03 | 1.08 | 1.13 | 1.18 | 1.23 | 1.28 | 0.5 |
| 0.86 | 0.89 | 0.94 | 0.99 | 1.04 | 1.09 | 1.14 | 1.19 | 1.24 | 1.29 | 0.5 |
| 0.87 | 0.90 | 0.95 | 1.00 | 1.05 | 1.10 | 1.15 | 1.20 | 1.25 | 1.30 | 0.5 |
| 0.88 | 0.91 | 0.96 | 1.01 | 1.06 | 1.11 | 1.16 | 1.21 | 1.26 | 1.31 | 0.5 |
| 0.89 | 0.92 | 0.97 | 1.02 | 1.07 | 1.12 | 1.17 | 1.22 | 1.27 | 1.32 | 0.5 |
| 0.90 | 0.93 | 0.98 | 1.03 | 1.08 | 1.13 | 1.18 | 1.23 | 1.28 | 1.33 | 0.5 |
| 0.91 | 0.94 | 0.99 | 1.04 | 1.09 | 1.14 | 1.19 | 1.24 | 1.29 | 1.34 | 0.5 |
| 0.92 | 0.95 | 1.00 | 1.05 | 1.10 | 1.15 | 1.20 | 1.25 | 1.30 | 1.35 | 0.5 |
| 0.93 | 0.96 | 1.01 | 1.06 | 1.11 | 1.16 | 1.21 | 1.26 | 1.31 | 1.36 | 0.5 |
| 0.94 | 0.97 | 1.02 | 1.07 | 1.12 | 1.17 | 1.22 | 1.27 | 1.32 | 1.37 | 0.5 |
| 0.95 | 0.98 | 1.03 | 1.08 | 1.13 | 1.18 | 1.23 | 1.28 | 1.33 | 1.38 | 0.5 |
| 0.96 | 0.99 | 1.04 | 1.09 | 1.14 | 1.19 | 1.24 | 1.29 | 1.34 | 1.39 | 0.5 |
| 0.97 | 1.00 | 1.05 | 1.10 | 1.15 | 1.20 | 1.25 | 1.30 | 1.35 | 1.40 | 0.5 |
| 0.98 | 1.01 | 1.06 | 1.11 | 1.16 | 1.21 | 1.26 | 1.31 | 1.36 | 1.41 | 0.5 |
| 0.99 | 1.02 | 1.07 | 1.12 | 1.17 | 1.22 | 1.27 | 1.32 | 1.37 | 1.42 | 0.5 |
| 1.00 | 1.03 | 1.08 | 1.13 | 1.18 | 1.23 | 1.28 | 1.33 | 1.38 | 1.43 | 0.5 |



参考文献

说明：参考文献中涉及部分网络资源，链接有可能发生变化。

中文书目

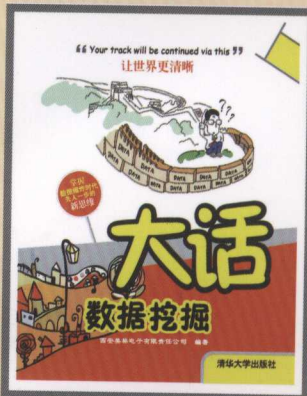
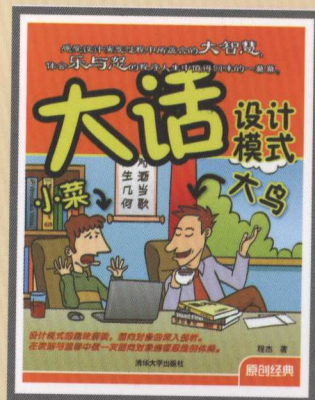
- [1] 陈文贤. 统计学(计算机应用)[M]. 中国台湾: 三民书局, 1992.
- [2] 陈文贤, 陈静枝. 统计学-中文统计 3.0 [M]. 中国台湾: 东华书局, 2012.
- [3] 桑慧敏. 机率与推论统计原理 [M]. 中国台湾: 美商麦格罗希尔国际股份有限公司, 2008.
- [4] 童甲春. 最新统计学全书研究所高普考各类考试复习自修范典 [M]. 中国台湾: 商务印书馆, 1989.
- [5] 颜月珠. 实用统计方法 [M]. 中国台湾: 三民书局, 1988.
- [6] 赵民德, 李纪难. 统计学 [M]. 中国台湾: 东华书局, 2005.
- [7] 赵民德. 赵家酒店 昔年种柳. <http://www.jds-online.com/blog/1999/02/02/>, 1999.
- [8] Moore DS. 统计学的世界 [M]. 郑惟厚, 译. 中国台湾: 天下远见出版股份有限公司, 2002.
- [9] 任立中, 周建亨, 陈静怡, 等. 统计学 [M]. 中国台湾: 前程文化事业公司, 2012.
- [10] 金勇进. 统计学: 第2版 [M]. 北京: 中国人民大学出版社, 2010.
- [11] 贾俊平, 何晓群, 金勇进. 统计学: 第5版 [M]. 北京: 中国人民大学出版社, 2012.
- [12] 贾俊平. 统计学基础 [M]. 北京: 中国人民大学出版社, 2010.
- [13] 刘树, 赵玉莲, 姜燕. 统计学 [M]. 北京: 清华大学出版社, 2010.
- [14] 车品觉. 决战大数据 [M]. 杭州: 浙江人民出版社, 2014.
- [15] 孙炎, 朱蔚青, 张银平. 统计基础与实务 [M]. 北京: 中国水利水电出版社, 2011.
- [16] 金秀, 于春海. 统计学 [M]. 北京: 清华大学出版社, 2014.
- [17] 徐晓岭, 王磊. 统计学 [M]. 北京: 人民邮电出版社, 2015.
- [18] 颜国勇. 机率论. 修订3版(网络版) [M]. 台南: 成功大学数学系, 2011.
- [19] 网站数据分析. 电子商务网站用户分析. <http://webdataanalysis.net/web-quantitative->

analysis/e-commerce-user-analysis/

- [20] 地理教室, 无国界. 第4课人口与都市 lovegeo.blogspot.com/2014/10/4_29.html
- [21] 王雪秋, 董小刚. 统计学理论与实务 [M]. 北京: 北京大学出版社, 2015.
- [22] 卢志飞, 孙忠宝. 应用统计学 [M]. 北京: 清华大学出版社, 2015.
- [23] 卢小广, 刘元欣. 统计学教程: 第2版 [M]. 北京: 清华大学出版社, 北京交通大学出版社, 2009.
- [24] International Labor Organization. 消费者价格指数手册理论与实践 [M]. 国际货币基金组织, 译. 北京: 中国财政经济出版社, 2008.

英文书目

- [1] Aczel AD, Sounderpandian JS. Complete Business Statistics [M]. 7th edition, 2009.
- [2] Akem A, Opryshko A. Why 'Lowe' when 'Young' and 'Laspeyres' are available?, Group of Experts on Consumer Price Indices. Reference paper 2 for the Higher-level Indices Workshop, United Nations Economic Commission for Europe, 2014.
- [3] Anderson DR, Sweeney DJ, Williams TA, Chen J. Statistics for Business and Economics [M]. 2006.
- [4] Armknecht P, Silver M. Post-Laspeyres: The Case for a New Formula for Compiling Consumer Price Indexes, IMF Working Paper, 2012
- [5] Baesens B. Analytics in a Big Data World [M]. 2014.
- [6] Black K. Business Statistics [M]. 4th edition, 2006.
- [7] Berenson ML, Levine DM, Krehbiel TC. Basic Business Statistics [M]. 10th edition, 2006.
- [8] Cox CP. A Handbook of Introductory Statistical Methods [M]. 1987.
- [9] Dowdy S, Wearden S. Statistics for Research [M]. 1983.
- [10] Freund JE, Williams W, Perles BM. Elementary Business Statistics [M]. 1988.
- [11] Huang N, Wimalaratne W, Pollard B. Choice of index number formula and the upper-level substitution bias in the Canadian CPI [M]. 2015.
- [12] Keller G. Statistics for Management and Economics [M]. 8th edition, 2009.
- [13] Kvanli AH, Guynes CS, Pavur RJ. Introduction to Business Statistics [M]. 3rd edition, 1992.



- 一图值千言，本书将统计学的观念、步骤、分类、关联等，尽量以图形表达，表达方式有：层次结构图、流程图、思维导图，以及因果表、比较表、决策法则表等，所以本书为统计学的学习地图。
- 本书配套软件《中文统计》几乎包括所有初等统计学的功能，在Excel（2003-2016版本适用）环境下，安装一个加载项，输入统计数据，就可以计算统计分析的结果。《中文统计》可以免费下载，仅提供给合法取得本书之读者使用。
- 每章例题、习题，补充教材也在网络资源下载。

 下载地址见本书前言

清华大学出版社数字出版网站

WQBook  书文局泉
www.wqbook.com

上架提示 统计学/大数据/算法

ISBN 978-7-302-45018-4



9 787302 450184 >

定价:59.00元